Supplementary material to the paper The VoicePrivacy 2020 Challenge: Results and findings

Natalia Tomashenko, Xin Wang, Emmanuel Vincent, Jose Patino, Brij Mohan Lal Srivastava, Paul-Gauthier Noé, Andreas Nautsch, Nicholas Evans, Junichi Yamagishi, Benjamin O'Brien, et al.

Supplementary material to the paper
The VoicePrivacy 2020 Challenge: Results and findings

Natalia Tomashenko[a,*], Xin Wang[b], Emmanuel Vincent[c], Jose Patino[d], Brij Mohan Lal Srivastava[f], Paul-Gauthier Noé[a], Andreas Nautsch[d], Nicholas Evans[d], Junichi Yamagishi[b,e], Benjamin O'Brien[g], Anaïs Chanclu[a], Jean-François Bonastre[a], Massimiliano Todisco[d], Mohamed Maouche[f]

[a] *LIA, University of Avignon, Avignon, France*
[b] *National Institute of Informatics (NII), Tokyo, Japan*
[c] *Université de Lorraine, CNRS, Inria, LORIA, France*
[d] *Audio Security and Privacy Group, EURECOM, France*
[e] *University of Edinburgh, UK*
[f] *Inria, France*
[g] *LPL, Aix-Marseille University, France*

Abstract

The VoicePrivacy 2020 Challenge focuses on developing anonymization solutions for speech technology. This report complements the summary results and analyses presented by Tomashenko et al. (2021). After quickly recalling the challenge design and the submitted anonymization systems, we provide more detailed results and analyses. First, we present objective evaluation results for the primary challenge metrics and for alternative metrics and attack models, and we compare them with each other. Second, we present subjective evaluation results for speaker verifiability, speech naturalness, and speech intelligibility. Finally, we compare these objective and subjective evaluation results with each other.

Keywords: privacy, anonymization, speech synthesis, voice conversion, speaker verification, automatic speech recognition, attack model, metrics, utility

Contents

*Corresponding author
Email address: natalia.tomashenko@univ-avignon.fr (Natalia Tomashenko)

1. Challenge design (summary)

Privacy preservation is formulated as a game between *users* who publish some data and *attackers* who access this data or data derived from it and wish to infer information about the users (Tomashenko et al., 2020b; Qian et al., 2018; Srivastava et al., 2020b).

Users (speakers) want to hide their identity while allowing all other downstream goals to be achieved. Attackers want to identify the speakers from one or more utterances.

The task of challenge participants is to develop an anonymization system. It should: (a) output a speech waveform, (b) hide speaker identity, (c) leave other speech characteristics unchanged, (d) ensure that all trial utterances from a given speaker are uttered by the same pseudo-speaker, while trial utterances from different speakers are uttered by different pseudo-speakers.

We consider objective and subjective privacy metrics to assess speaker re-identification and linkability. We also propose objective and subjective utility metrics in order to assess the fulfillment of the user goals. Specifically, we consider ASR performance using a model trained on clean and anonymized data, as well as subjective speech intelligibility and naturalness.

For objective evaluation of anonymization performance, two systems were trained to assess the following characteristics: speaker verifiability and ability of the anonymization system to preserve linguistic information in the anonymized speech. The first system, denoted ASV_{eval}, is an automatic speaker verification (ASV) system. The second system, denoted ASR_{eval}, is an automatic speech recognition (ASR) system. These two systems were used in the VoicePrivacy official challenge setup (Tomashenko et al., 2020b,a). In addition, we trained ASV and ASR systems on anonymized speech data: $ASV_{\text{eval}}^{\text{anon}}$ and $ASR_{\text{eval}}^{\text{anon}}$.

The **objective evaluation** metrics for privacy and utility include:

1. Equal error rate (EER)
2. Log-likelihood-ratio cost function (C_{llr} and $C_{\text{llr}}^{\text{min}}$)
3. Metrics computed from voice similarity matrices: de-identification and voice distinctiveness preservation
4. Linkability
5. Zero evidence biometric recognition assessment (ZEBRA) framework metrics: expected privacy disclosure (population) and worst case privacy disclosure (individual)
6. Word error rate (WER).

Metrics #1–5 were estimated using anonymized trial data, original or anonymized enrollment data, and ASV_{eval} or $ASV_{\text{eval}}^{\text{anon}}$ models in different conditions corresponding to different attack models with increasing strength: *ignorant*, *lazy-informed*, and *semi-informed*. The WER was computed on original or anonymized trial data using ASR_{eval} or $ASR_{\text{eval}}^{\text{anon}}$.

We consider the following **subjective evaluation** metrics:

1. Speaker verifiability
2. Speaker linkability
3. Speech intelligibility
4. Speech naturalness.

2. Anonymization systems (summary)

Two different anonymization systems were provided as challenge baselines[1]:

- **B1** (primary baseline): extraction of pitch (F0) and bottleneck (BN) features followed by speech synthesis (SS) using an anonymized x-vector, an SS acoustic model (AM) and a neural source-filter (NSF) model (Tomashenko et al., 2020b; Srivastava et al., 2020a);
- **B2** (secondary baseline): anonymization using McAdams coefficient (Patino et al., 2021).

Table 1 provides an overview of the systems submitted by the challenge participants. Most systems were inspired by the primary baseline, one system was based upon the secondary baseline, and two systems are not related to either.

Table 1: Challenge submissions, team names and organizations. Submission identifiers (IDs) for each system are shown in the last column (ID) and comprise: <team id: first letter of the team name><submission deadline[2]: 1 or 2><c, if the system is contrastive><index of the contrastive system>. Blue star symbols ⋆ in the first column indicate teams submitted the anonymized training data for post-evaluation analysis; (1) and (2) – teams developed their systems from the baseline-1 and baseline-2 respectively, and ◯ – other submissions.

Team (Reference)	System (Details)
AIS-lab JAIST (Mawalim et al., 2020) (A)	**A1** (x-vector anonymization using statistical-based ensemble regression modeling) **A2** (using singular value modification)
DA-IICT Speech Group (Gupta et al., 2020) (D)	**D1** (modifications to pole radius)
Idiap-NKI (Dubagunta et al., 2020) (I)	**I1** (modifications to formants, F0 and speaking rate)
Kyoto Team (Han et al., 2020) (K) ⋆	**K2** (anonymization using x-vectors, SS models and a voice-indistinguishability metric)
MultiSpeech (Champion et al., 2020) (M) ⋆	**M1** (end-to-end ASR model for BN feature extraction) **M1c1** (semi-adversarial training to learn linguistic features while masking speaker information) **M1c2** (copy-synthesis (original x-vectors)) **M1c3** (x-vectors provided to SS AM are anonymized, x-vectors provided to NSF are original) **M1c4** (x-vectors provided to SS AM are original, x-vectors provided to NSF are anonymized)
Oxford System Security Lab (Turner et al., 2020) (O) ⋆	**O1** (keeping original distribution of cosine distances between speaker x-vectors; GMM for sampling x-vectors in a PCA-reduced space followed by projection to the original dimension) **O1c1** (**O1** with forced dissimilarity between original and generated x-vectors)
Sigma Technologies SLU (Espinoza-Cuadros et al., 2020) (S) ⋆	**S1** (**S1c1** applied on the top of the **B1** x-vector anonymization) **S1c1** (domain-adversarial training; autoencoders: using gender, accent, speaker id outputs corresponding to adversarial branches in the neural network for x-vector reconstruction) **S2** (**S2c1** applied on the top of the **B1** x-vector anonymization) **S2c1** (**S1c1** with parameter optimization)
PingAn (Huang, 2020)	a non-challenge entry work; this team worked on the development of stronger attack models for ASV evaluation.

3. Objective evaluation

This section presents objective evaluation results for the primary challenge metrics (Section 3.1), alternative metrics and attack models (Sections 3.2—3.6), as well as comparative results for different privacy metrics (Section 3.7).

[1]Baseline systems are available online: `https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2020`
[2]Deadline-1: 8th May 2020; deadline-2: 16th June 2020.

3.1. EER, C_{llr}, and C_{llr}^{min}

Equal error rate (EER). Denoting by $P_{\text{fa}}(\theta)$ and $P_{\text{miss}}(\theta)$ the false alarm and miss rates at threshold θ, the EER corresponds to the threshold θ_{EER} at which the two detection error rates are equal, i.e., EER = $P_{\text{fa}}(\theta_{\text{EER}}) = P_{\text{miss}}(\theta_{\text{EER}})$. We also considered ROC (receiver operating characteristic) convex hull (Provost & Fawcett, 2001) EER, or ROCCH-EER (Brummer, 2010) for experiments in Section 5. The ROCCH is obtained by interpolating between the points of the ROC curve.

Log-likelihood-ratio cost function (C_{llr} and C_{llr}^{min}). The log-likelihood-ratio cost function (C_{llr}) was proposed by Brümmer & Du Preez (2006) as an *application-independent* evaluation objective and is defined as follows:

$$C_{\text{llr}} = \frac{1}{2}\left(\frac{1}{N_{\text{tar}}}\sum_{i\in\text{tar}}\log_2\left(1+e^{-\text{LLR}_i}\right) + \frac{1}{N_{\text{imp}}}\sum_{j\in\text{imp}}\log_2\left(1+e^{\text{LLR}_j}\right)\right), \tag{1}$$

where N_{tar} and N_{imp} are respectively the number of target and impostor log-likelihood ratio (LLR) values in the evaluation set. C_{llr} can be decomposed into a discrimination loss ($C_{\text{llr}}^{\text{min}}$) and a calibration loss ($C_{\text{llr}} - C_{\text{llr}}^{\text{min}}$) (Brümmer & Du Preez, 2006). The $C_{\text{llr}}^{\text{min}}$ is estimated by optimal calibration using monotonic transformation of scores to their empirical LLR values. To obtain this monotonic transformation, the pool adjacent violators (PAV) to LLR algorithm is used (Brümmer & Du Preez, 2006; Ramos & Gonzalez-Rodriguez, 2008).

$C_{\text{llr}}^{\text{min}}$ relates to the EER through the receiver operating characteristic (ROC) convex hull: $C_{\text{llr}}^{\text{min}}$ is its scalar summary and the EER an extreme point (Brümmer & De Villiers, 2011). If the EER of the convex hull changes, the entire hull is affected due to convexity and hence is $C_{\text{llr}}^{\text{min}}$; by contrast, a change in $C_{\text{llr}}^{\text{min}}$ does not need to affect the EER.

Results. Tables 2 and 3 provide the privacy objective evaluation results in terms of EER, C_{llr}, and $C_{\text{llr}}^{\text{min}}$ for two attack models — *ignorant* (**oa**: original enrollment and anonymized trial data) and *lazy-informed* (**aa**: anonymized enrollment and anonymized trial data) — on the VoicePrivacy development and test datasets for all submitted and baseline anonymization systems. Figures 1 and 2 provide the summary of EER and C_{llr} on the test datasets for ignorant and lazy-informed attack models.

3.2. Zero evidence biometric recognition assessment (ZEBRA) framework

Expected and worst-case privacy disclosure. Expected and worst-case privacy disclosure metrics have been proposed by Nautsch et al. (2020) for the zero evidence biometric recognition assessment (ZEBRA) framework. They measure the average level of privacy preservation afforded by a given safeguard for a population and the worst-case privacy disclosure for an individual. If the expected privacy disclosure D_{ECE} is equal to 0, then we assume that perfect privacy (zero evidence) is achieved.

Results. Expected and worst-case privacy disclosure results are given in Table 4 (development) and Table 5 (test) for ignorant and lazy-informed attack models. ZEBRA assessment empirical cross entropy (ECE) plots are shown in Figures 3, 4 (*LibriSpeech*) and Figures 5, 6 (*VCTK different*) in the form: (expected privacy disclosure, worst-case privacy disclosure, categorical tags of worst-case privacy disclosure).[3]

[3]Categorical tags of worst-case privacy disclosure (Nautsch et al., 2020)

Tag	Posterior odds ratio (flat prior)
0	50 : 50 (flat posterior)
A	more disclosure than 50 : 50
B	one wrong in 10 to 100
C	one wrong in 100 to 10 000
D	one wrong in 10 000 to 100 000
E	one wrong in 100 000 to 1 000 000
F	one wrong in at least 1 000 000

Table 2: Objective results: EER, C_{llr}, and $C_{\text{llr}}^{\text{min}}$ for ignorant and lazy-informed attack models on the development data. Larger EER and $C_{\text{llr}}^{\text{min}}$ values correspond to better privacy.

EER – Ignorant (oa)																				
Data	**Gender**	**Orig**	**A1**	**A2**	**B1**	**B2**	**D1**	**I1**	**K2**	**M1**	**M1c1**	**M1c2**	**M1c3**	**M1c4**	**O1**	**O1c1**	**S2**	**S1**	**S1c1**	**S2c1**
LibriSpeech	female	8.66	46.88	51.85	50.14	35.23	33.10	31.25	43.32	50.99	50.71	25.57	50.43	26.42	43.32	43.89	45.88	43.32	43.47	44.32
	male	1.24	54.19	59.01	57.76	18.17	19.72	15.37	41.93	53.88	54.97	24.07	54.66	24.22	50.31	49.69	50.00	54.35	40.37	49.84
VCTK different	female	2.86	49.69	50.65	49.97	35.49	33.91	13.53	54.69	53.73	53.62	27.23	48.34	26.05	46.60	46.32	50.31	47.61	40.09	49.75
	male	1.44	52.75	55.88	53.95	28.34	26.20	26.30	44.96	50.62	51.17	18.51	54.29	18.86	45.11	45.66	45.21	48.73	39.90	44.71
VCTK common	female	2.62	48.55	50.58	49.71	34.01	32.56	18.60	47.38	50.58	51.16	29.94	48.84	30.23	44.77	46.80	50.58	47.09	45.06	50.58
	male	1.43	51.85	57.83	54.99	23.93	24.50	29.06	49.29	53.28	53.28	21.37	54.13	21.94	49.29	49.86	47.58	48.43	42.45	48.15
EER – Lazy-informed (aa)																				
Data	**Gender**	**Orig**	**A1**	**A2**	**B1**	**B2**	**D1**	**I1**	**K2**	**M1**	**M1c1**	**M1c2**	**M1c3**	**M1c4**	**O1**	**O1c1**	**S2**	**S1**	**S1c1**	**S2c1**
LibriSpeech	female	8.66	33.24	33.24	36.79	23.72	24.29	25.14	3.69	33.38	31.11	19.60	35.09	18.75	36.79	36.93	39.63	39.77	30.82	39.77
	male	1.24	32.76	28.88	34.16	10.87	11.02	18.63	2.17	29.66	28.57	17.86	32.14	16.30	41.61	41.61	44.25	39.60	35.09	43.63
VCTK different	female	2.86	26.90	26.61	26.11	15.83	14.04	15.67	4.32	23.19	24.26	11.01	23.86	11.45	31.05	30.66	31.22	24.42	17.46	30.38
	male	1.44	30.72	25.51	30.92	11.17	13.50	14.64	9.03	32.06	33.05	9.73	30.12	9.93	39.06	38.81	34.74	34.59	29.33	33.55
VCTK common	female	2.62	24.42	24.42	27.91	11.63	10.76	16.86	3.78	33.72	32.56	11.92	25.00	11.05	32.85	31.69	29.65	27.91	23.84	29.94
	male	1.43	31.05	26.50	33.33	10.54	12.54	20.23	4.84	38.18	38.18	9.12	31.91	8.83	42.74	41.31	39.32	39.32	30.77	38.18

C_{llr} – Ignorant (oa)																				
Data	**Gender**	**Orig**	**A1**	**A2**	**B1**	**B2**	**D1**	**I1**	**K2**	**M1**	**M1c1**	**M1c2**	**M1c3**	**M1c4**	**O1**	**O1c1**	**S2**	**S1**	**S1c1**	**S2c1**
LibriSpeech	female	42.9	164.2	147.2	144.1	116.8	115.5	115.3	148.6	150.3	150.2	114.3	144.6	110.8	134.7	134.3	171.3	144.6	141.0	171.5
	male	14.2	166.7	170.6	169.0	105.8	112.1	77.6	148.4	155.8	156.7	111.1	165.7	111.5	147.8	147.8	153.4	153.2	137.1	153.4
VCTK different	female	1.1	173.5	164.3	166.0	90.6	102.5	23.6	181.4	178.1	173.8	113.8	159.7	118.4	168.0	167.3	179.3	163.1	153.4	177.9
	male	1.2	162.1	166.5	167.5	98.5	101.2	75.1	138.6	161.9	162.7	104.9	163.8	108.1	154.8	155.2	152.0	153.2	144.5	151.3
VCTK common	female	0.9	182.5	167.5	172.0	85.9	100.4	28.3	159.6	173.7	168.9	102.0	162.1	109.6	181.1	180.8	183.9	165.8	163.3	183.3
	male	1.6	187.1	191.7	192.9	90.8	97.4	75.4	160.8	173.0	174.5	113.1	187.2	118.9	179.4	179.6	172.3	170.1	161.4	172.0
C_{llr} – Lazy-informed (aa)																				
Data	**Gender**	**Orig**	**A1**	**A2**	**B1**	**B2**	**D1**	**I1**	**K2**	**M1**	**M1c1**	**M1c2**	**M1c3**	**M1c4**	**O1**	**O1c1**	**S2**	**S1**	**S1c1**	**S2c1**
LibriSpeech	female	42.9	15.1	14.3	16.3	11.7	15.4	9.3	5.1	25.2	25.4	10.7	15.8	7.5	11.1	11.2	23.0	23.3	15.0	22.3
	male	14.2	21.1	18.4	24.7	11.9	15.9	15.7	1.7	37.9	34.2	7.6	22.8	7.9	12.1	11.9	38.4	35.2	27.4	36.9
VCTK different	female	1.1	11.1	8.7	8.4	39.9	44.3	6.3	2.4	21.6	22.6	3.1	7.5	3.1	13.5	12.9	15.7	9.8	5.8	14.9
	male	1.2	20.0	18.3	23.8	23.2	36.6	3.8	7.0	46.5	40.1	12.2	23.5	11.6	10.4	10.6	31.0	38.2	29.0	30.4
VCTK common	female	0.9	8.6	7.1	7.2	43.6	43.8	11.1	4.5	22.2	21.9	4.1	6.8	3.8	11.5	11.6	15.1	9.8	8.0	14.3
	male	1.6	18.5	18.2	23.9	25.0	34.2	7.6	5.4	39.1	37.9	8.7	23.4	8.3	11.8	10.9	31.7	35.7	23.7	30.4

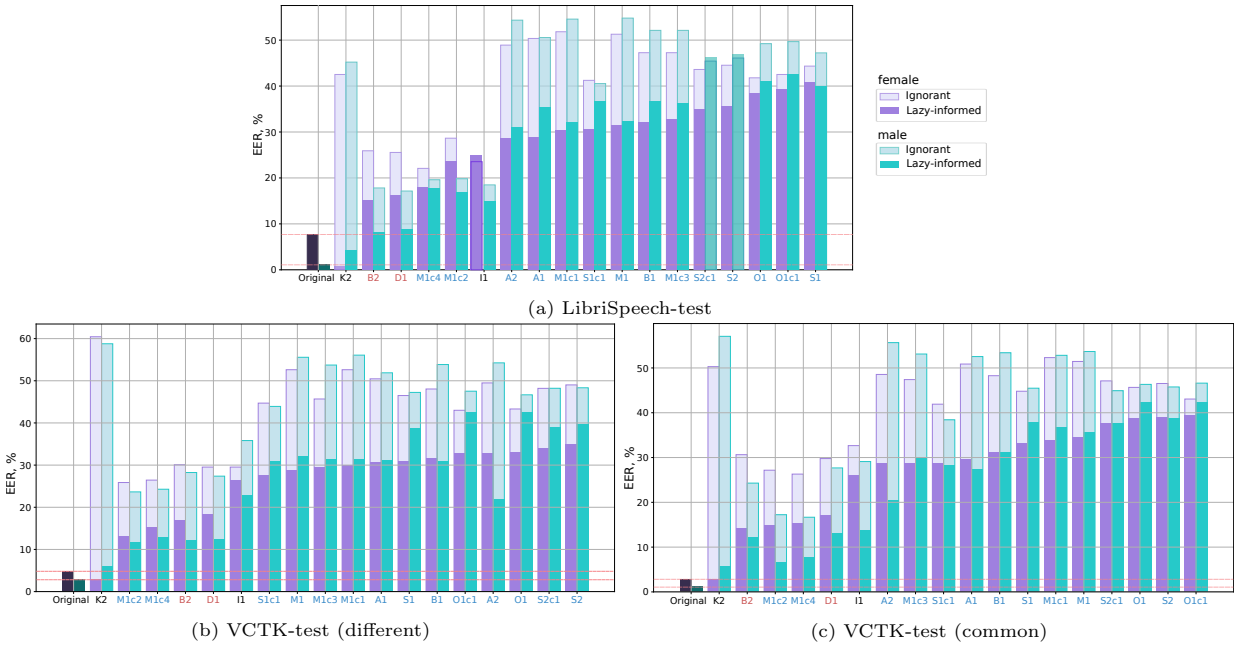
$C_{\text{llr}}^{\text{min}}$ – Ignorant (oa)																				
Data	**Gender**	**Orig**	**A1**	**A2**	**B1**	**B2**	**D1**	**I1**	**K2**	**M1**	**M1c1**	**M1c2**	**M1c3**	**M1c4**	**O1**	**O1c1**	**S2**	**S1**	**S1c1**	**S2c1**
LibriSpeech	female	0.304	0.990	0.998	0.996	0.820	0.808	0.812	0.944	0.991	0.991	0.726	0.997	0.712	0.970	0.968	0.981	0.970	0.953	0.975
	male	0.034	1.000	0.997	0.999	0.527	0.580	0.457	0.914	0.978	0.977	0.697	0.998	0.689	0.977	0.978	0.991	0.996	0.917	0.989
VCTK different	female	0.100	0.969	0.988	0.989	0.907	0.898	0.443	1.000	0.996	0.993	0.748	0.985	0.731	0.945	0.944	0.963	0.970	0.893	0.958
	male	0.052	1.000	1.000	1.000	0.743	0.731	0.712	0.978	0.999	0.997	0.613	1.000	0.630	0.990	0.992	0.991	0.998	0.958	0.988
VCTK common	female	0.088	0.990	0.996	0.995	0.877	0.864	0.553	0.966	0.994	0.994	0.773	0.988	0.771	0.957	0.959	0.982	0.975	0.946	0.976
	male	0.050	0.997	1.000	0.999	0.671	0.672	0.703	0.991	0.993	0.995	0.598	0.999	0.620	0.990	0.991	0.987	0.993	0.953	0.985
$C_{\text{llr}}^{\text{min}}$ – Lazy-informed (aa)																				
Data	**Gender**	**Orig**	**A1**	**A2**	**B1**	**B2**	**D1**	**I1**	**K2**	**M1**	**M1c1**	**M1c2**	**M1c3**	**M1c4**	**O1**	**O1c1**	**S2**	**S1**	**S1c1**	**S2c1**
LibriSpeech	female	0.304	0.872	0.864	0.894	0.621	0.650	0.663	0.138	0.833	0.824	0.566	0.826	0.547	0.898	0.902	0.924	0.922	0.800	0.921
	male	0.034	0.854	0.780	0.867	0.358	0.370	0.559	0.086	0.799	0.787	0.516	0.828	0.508	0.947	0.944	0.969	0.933	0.844	0.964
VCTK different	female	0.100	0.771	0.770	0.760	0.503	0.452	0.505	0.166	0.696	0.722	0.376	0.708	0.376	0.838	0.826	0.834	0.713	0.503	0.829
	male	0.052	0.836	0.738	0.839	0.385	0.435	0.388	0.304	0.858	0.876	0.346	0.823	0.347	0.938	0.936	0.898	0.887	0.806	0.881
VCTK common	female	0.088	0.696	0.705	0.741	0.366	0.353	0.509	0.140	0.869	0.855	0.371	0.712	0.365	0.843	0.832	0.814	0.775	0.641	0.802
	male	0.050	0.816	0.704	0.840	0.316	0.394	0.563	0.185	0.929	0.921	0.316	0.820	0.318	0.957	0.950	0.936	0.931	0.787	0.921

Table 3: Objective results: EER, C_{llr}, and $C_{\text{llr}}^{\text{min}}$ for ignorant and lazy-informed attack models on the test data. Larger EER and $C_{\text{llr}}^{\text{min}}$ values correspond to better privacy.

EER – Ignorant (oa)																				
Data	**Gender**	**Orig**	**A1**	**A2**	**B1**	**B2**	**D1**	**I1**	**K2**	**M1**	**M1c1**	**M1c2**	**M1c3**	**M1c4**	**O1**	**O1c1**	**S2**	**S1**	**S1c1**	**S2c1**
LibriSpeech	female	7.66	50.36	48.91	47.26	25.91	25.55	23.54	42.52	51.28	51.82	28.65	47.26	22.08	41.79	42.52	44.53	44.34	41.24	43.61
	male	1.11	50.56	54.34	52.12	17.82	17.15	18.49	45.21	54.79	54.57	19.82	52.12	19.6	49.22	49.67	46.1	47.22	40.53	45.43
VCTK different	female	4.89	50.46	49.49	48.05	30.09	29.53	29.53	60.44	52.62	52.62	25.87	45.68	26.44	43.31	43.0	49.02	46.5	44.7	48.2
	male	2.07	51.89	54.25	53.85	28.24	27.38	35.82	58.78	55.57	56.08	23.65	53.73	24.28	46.67	47.53	48.34	47.24	43.92	48.22
VCTK common	female	2.89	50.87	48.55	48.27	30.64	29.77	32.66	50.29	51.45	52.31	27.17	47.4	26.3	45.66	43.06	46.53	44.8	41.91	47.11
	male	1.13	52.54	55.65	53.39	24.29	27.68	29.1	57.06	53.67	52.82	17.23	53.11	16.67	46.33	46.61	45.76	45.48	38.42	44.92
EER – Lazy-informed (aa)																				
Data	**Gender**	**Orig**	**A1**	**A2**	**B1**	**B2**	**D1**	**I1**	**K2**	**M1**	**M1c1**	**M1c2**	**M1c3**	**M1c4**	**O1**	**O1c1**	**S2**	**S1**	**S1c1**	**S2c1**
LibriSpeech	female	7.66	28.83	28.65	32.12	15.15	16.24	24.82	0.73	31.39	30.47	23.54	32.66	17.88	38.32	39.23	35.58	40.88	30.66	35.04
	male	1.11	35.41	30.96	36.75	8.24	8.91	14.92	4.23	32.29	32.07	16.93	36.30	17.82	40.98	42.54	46.99	39.87	36.75	46.33
VCTK different	female	4.89	30.81	32.92	31.74	16.92	18.42	26.34	3.04	28.81	29.94	13.07	29.63	15.23	32.97	32.82	35.08	30.97	27.73	34.10
	male	2.07	31.11	21.87	30.94	12.23	12.51	22.96	5.97	32.20	31.52	11.77	31.46	12.92	42.65	42.48	39.61	38.69	31.00	38.98
VCTK common	female	2.89	29.48	28.61	31.21	14.16	17.05	26.01	2.89	34.39	33.82	14.74	28.61	15.32	38.73	39.31	39.02	33.24	28.61	37.57
	male	1.13	27.40	20.34	31.07	12.15	12.99	13.84	5.65	35.59	36.72	6.50	29.94	7.63	42.37	42.37	38.70	37.85	28.25	37.57

C_{llr} – Ignorant (oa)																				
Data	**Gender**	**Orig**	**A1**	**A2**	**B1**	**B2**	**D1**	**I1**	**K2**	**M1**	**M1c1**	**M1c2**	**M1c3**	**M1c4**	**O1**	**O1c1**	**S2**	**S1**	**S1c1**	**S2c1**
LibriSpeech	female	26.8	165.1	151.9	151.8	115.5	119.5	94.6	155.7	156.0	155.5	112.3	147.8	110.9	145.5	146.8	166.8	144.0	143.7	168.5
	male	15.3	164.4	168.8	166.7	106.4	110.9	89.0	156.9	159.0	160.1	117.6	165.8	117.6	174.1	175.1	155.7	151.7	146.1	155.5
VCTK different	female	1.5	154.5	142.9	146.9	93.2	103.7	41.0	171.6	157.0	152.3	97.4	141.5	104.3	148.3	147.8	156.7	146.1	142.8	156.0
	male	1.8	163.7	164.8	167.8	101.6	111.9	79.4	162.6	166.0	169.2	111.3	165.9	111.6	162.5	162.6	157.6	156.2	152.2	157.5
VCTK common	female	0.9	170.4	157.7	162.5	94.0	107.9	51.9	170.6	176.8	171.9	91.6	155.7	99.6	161.6	160.7	172.0	157.7	152.9	171.7
	male	1.0	184.3	186.5	190.1	99.3	107.5	68.1	156.2	170.7	172.5	116.5	187.5	118.4	184.6	185.5	172.5	168.5	161.3	172.3
C_{llr} – Lazy-informed (aa)																				
Data	**Gender**	**Orig**	**A1**	**A2**	**B1**	**B2**	**D1**	**I1**	**K2**	**M1**	**M1c1**	**M1c2**	**M1c3**	**M1c4**	**O1**	**O1c1**	**S2**	**S1**	**S1c1**	**S2c1**
LibriSpeech	female	26.8	13.6	12.7	16.3	12.6	15.2	10.2	2.4	29.8	29.4	18.2	23.7	6.4	13.4	13.9	27.6	29.3	13.4	27.1
	male	15.3	28.0	24.2	33.9	15.4	21.9	10.7	4.8	41.5	38.3	10.3	33.4	10.6	17.6	19.2	48.3	44.8	39.2	47.7
VCTK different	female	1.5	13.6	11.3	11.5	41.3	49.7	11.9	1.6	30.3	30.6	6.5	10.1	7.1	17.8	18.1	21.9	15.1	12.2	21.3
	male	1.8	19.5	13.3	23.8	25.1	35.2	7.6	4.3	45.6	39.5	11.1	23.0	12.0	14.4	14.7	33.9	41.3	30.4	33.5
VCTK common	female	0.9	10.2	8.8	9.0	42.7	47.4	13.2	3.0	24.1	23.9	6.2	8.1	6.1	11.4	11.5	17.7	12.4	10.0	17.3
	male	1.0	15.6	9.8	21.7	28.2	36.1	5.3	7.4	34.5	33.1	6.8	21.0	7.3	10.4	10.7	32.1	34.0	21.9	30.6

$C_{\text{llr}}^{\text{min}}$ – Ignorant (oa)																				
Data	**Gender**	**Orig**	**A1**	**A2**	**B1**	**B2**	**D1**	**I1**	**K2**	**M1**	**M1c1**	**M1c2**	**M1c3**	**M1c4**	**O1**	**O1c1**	**S2**	**S1**	**S1c1**	**S2c1**
LibriSpeech	female	0.183	0.994	0.996	0.995	0.684	0.692	0.618	0.936	0.996	0.997	0.697	0.987	0.632	0.969	0.973	0.980	0.975	0.955	0.972
	male	0.041	0.997	1.000	0.999	0.499	0.491	0.498	0.898	0.997	0.999	0.608	0.999	0.601	0.983	0.982	0.985	0.983	0.935	0.981
VCTK different	female	0.169	1.000	0.999	0.998	0.794	0.798	0.742	1.000	1.000	0.999	0.719	0.993	0.741	0.981	0.978	0.996	0.992	0.984	0.996
	male	0.072	1.000	1.000	1.000	0.720	0.729	0.853	0.989	1.000	1.000	0.687	1.000	0.680	0.992	0.991	0.996	0.996	0.981	0.996
VCTK common	female	0.091	0.989	0.991	0.994	0.808	0.799	0.770	0.996	0.999	0.998	0.733	0.989	0.738	0.976	0.978	0.982	0.976	0.946	0.982
	male	0.036	0.999	1.000	1.000	0.713	0.720	0.699	0.973	0.998	0.998	0.514	1.000	0.508	0.988	0.987	0.987	0.990	0.941	0.985
$C_{\text{llr}}^{\text{min}}$ – Lazy-informed (aa)																				
Data	**Gender**	**Orig**	**A1**	**A2**	**B1**	**B2**	**D1**	**I1**	**K2**	**M1**	**M1c1**	**M1c2**	**M1c3**	**M1c4**	**O1**	**O1c1**	**S2**	**S1**	**S1c1**	**S2c1**
LibriSpeech	female	0.183	0.765	0.777	0.839	0.491	0.512	0.592	0.025	0.838	0.813	0.639	0.843	0.532	0.915	0.928	0.897	0.901	0.792	0.886
	male	0.041	0.878	0.806	0.903	0.264	0.279	0.434	0.133	0.840	0.821	0.521	0.898	0.549	0.960	0.966	0.979	0.939	0.910	0.978
VCTK different	female	0.169	0.842	0.871	0.847	0.547	0.580	0.752	0.113	0.810	0.823	0.443	0.803	0.487	0.861	0.859	0.896	0.836	0.781	0.882
	male	0.072	0.849	0.666	0.834	0.398	0.424	0.666	0.226	0.863	0.863	0.396	0.841	0.431	0.968	0.965	0.954	0.943	0.833	0.947
VCTK common	female	0.091	0.783	0.800	0.830	0.464	0.500	0.698	0.095	0.887	0.877	0.438	0.775	0.458	0.910	0.903	0.919	0.875	0.792	0.913
	male	0.036	0.769	0.614	0.835	0.354	0.388	0.453	0.202	0.904	0.883	0.218	0.815	0.245	0.957	0.960	0.943	0.923	0.800	0.927

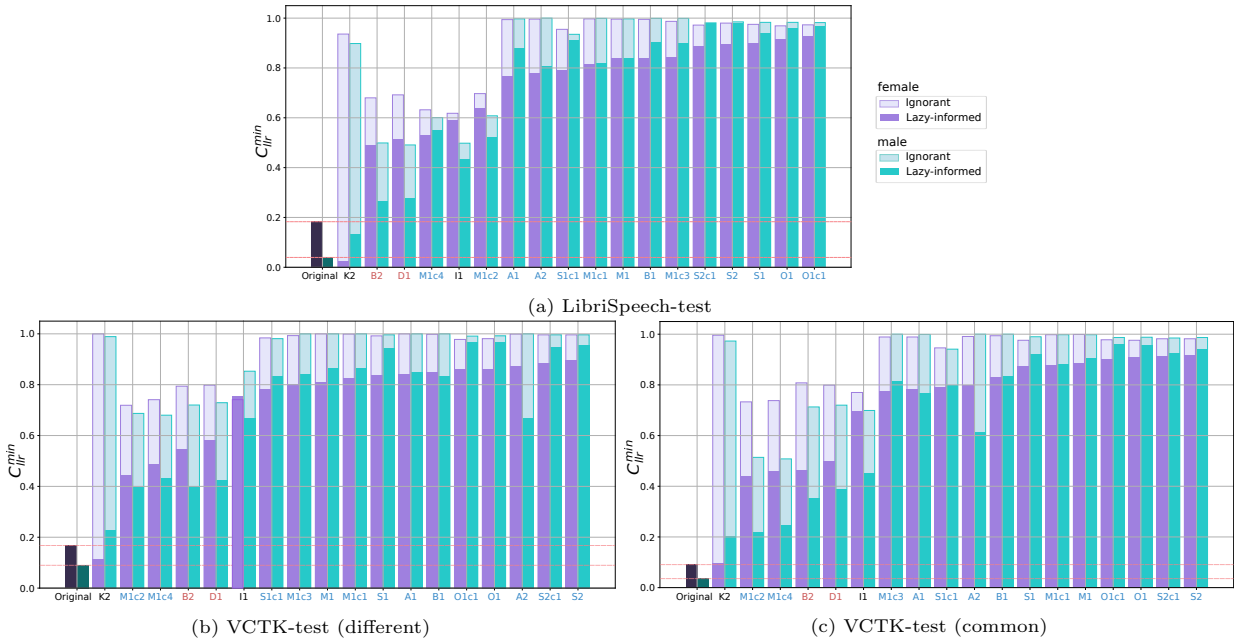


(a) LibriSpeech-test

(b) VCTK-test (different)

(c) VCTK-test (common)

Figure 1: EER results on the test datasets for different anonymization systems and for original data. Blue and red colors in the system notations indicate systems developed from **B1** and **B2** respectively. The results are ordered by EER values on female speakers for the lazy-informed attack model. Larger EER values correspond to better privacy.

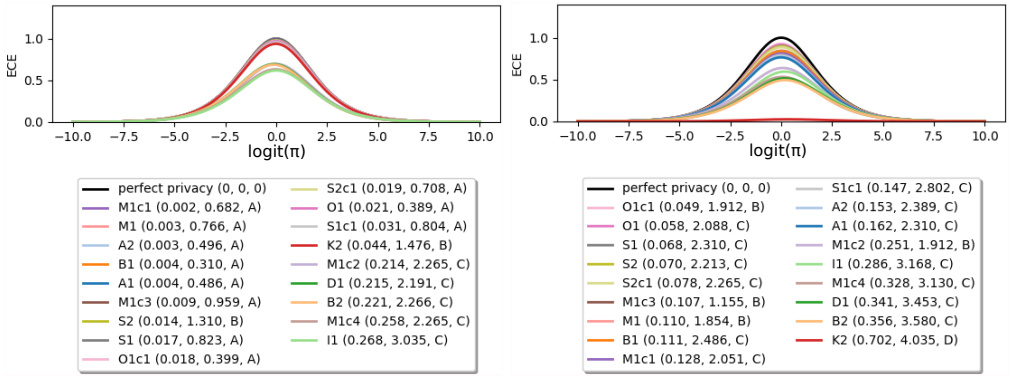


(a) LibriSpeech-test

(b) VCTK-test (different)

(c) VCTK-test (common)

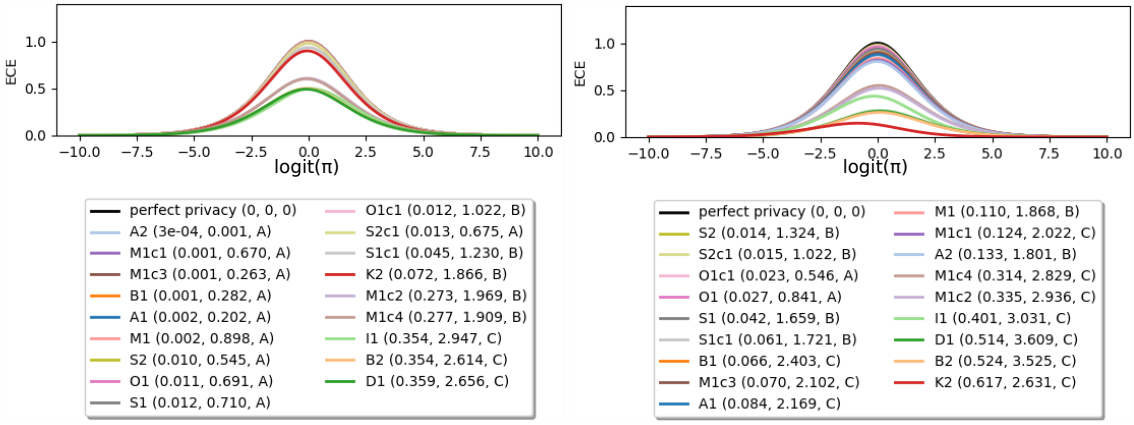
Figure 2: $C_{\text{llr}}^{\text{min}}$ results on the test datasets for different anonymization systems and for original data; **oa**: original enrollment and anonymized trial data; and **aa**: both enrolment and trial data are anonymized. Blue and red colors in the system notations indicate systems developed from **B1** and **B2** respectively. The results are ordered by $C_{\text{llr}}^{\text{min}}$ values on female speakers for the lazy-informed attack model. Larger $C_{\text{llr}}^{\text{min}}$ values correspond to better privacy.



(a) Ignorant attack model (oa)

(b) Lazy-informed attack model (aa)

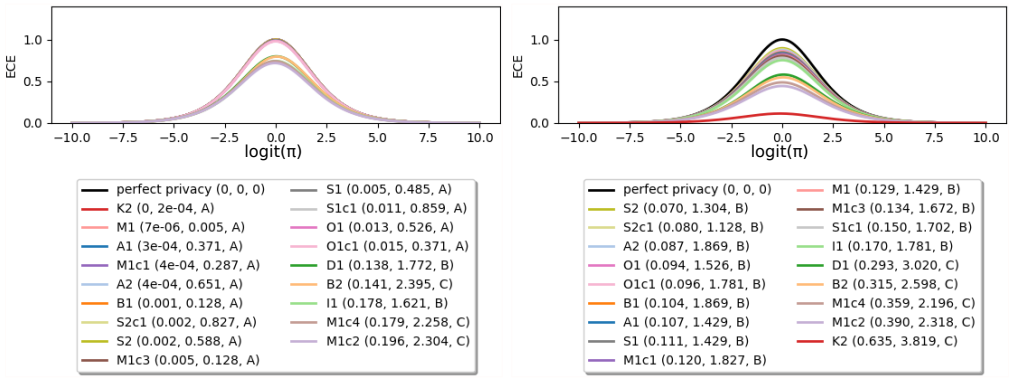
Figure 3: ZEBRA assessment with ECE profiles on *LibriSpeech-test* for female speakers.



(a) Ignorant attack model (oa)

(b) Lazy-informed attack model (aa)

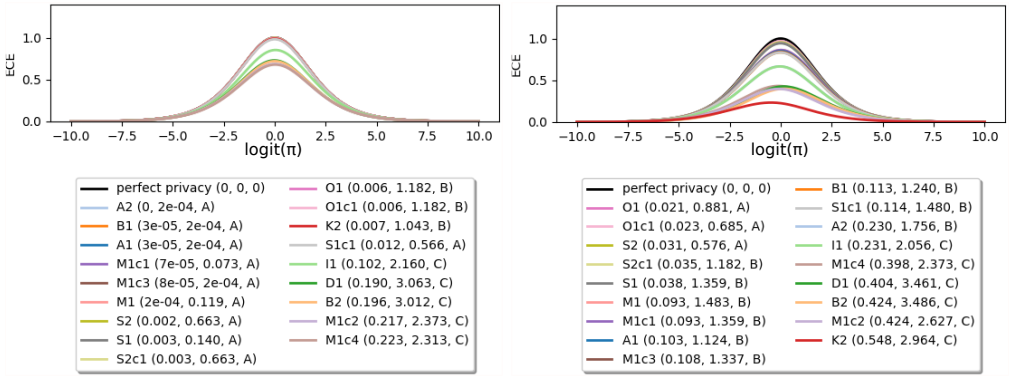
Figure 4: ZEBRA assessment with ECE profiles on *LibriSpeech-test* for male speakers.



(a) Ignorant attack model (oa)

(b) Lazy-informed attack model (aa)

Figure 5: ZEBRA assessment with ECE profiles on *VCTK-test (different)* for female speakers.



(a) Ignorant attack model (oa)

(b) Lazy-informed attack model (aa)

Figure 6: ZEBRA assessment with ECE profiles on *VCTK-test (different)* for male speakers.

Table 4: Objective results: ZEBRA expected and worst-case privacy disclosure for ignorant and lazy-informed attack models on the development data.

ZEBRA: Expected privacy disclosure (population), bit – Ignorant (oa)																				
Data	**Gender**	**Orig**	**A1**	**A2**	**B1**	**B2**	**D1**	**I1**	**K2**	**M1**	**M1c1**	**M1c2**	**M1c3**	**M1c4**	**O1**	**O1c1**	**S2**	**S1**	**S1c1**	**S2c1**
LibriSpeech	female	0.492	0.007	0.002	0.002	0.126	0.135	0.129	0.038	0.006	0.006	0.190	0.002	0.201	0.020	0.022	0.013	0.020	0.033	0.017
	male	0.696	0.000	0.002	0.001	0.334	0.295	0.383	0.060	0.016	0.016	0.211	0.001	0.216	0.016	0.015	0.006	0.002	0.058	0.008
VCTK different	female	0.646	0.021	0.008	0.008	0.063	0.069	0.391	0.000	0.003	0.005	0.174	0.010	0.186	0.038	0.039	0.025	0.021	0.075	0.029
	male	0.682	0.000	0.000	0.000	0.178	0.186	0.200	0.015	0.001	0.002	0.267	0.000	0.254	0.006	0.005	0.006	0.001	0.028	0.008
VCTK common	female	0.653	0.007	0.003	0.004	0.083	0.093	0.311	0.023	0.004	0.004	0.159	0.008	0.160	0.030	0.029	0.013	0.017	0.038	0.017
	male	0.683	0.002	0.000	0.001	0.228	0.229	0.207	0.006	0.005	0.004	0.282	0.001	0.266	0.007	0.006	0.009	0.005	0.032	0.010
ZEBRA: Expected privacy disclosure (population), bit – Lazy-informed (aa)																				
Data	**Gender**	**Orig**	**A1**	**A2**	**B1**	**B2**	**D1**	**I1**	**K2**	**M1**	**M1c1**	**M1c2**	**M1c3**	**M1c4**	**O1**	**O1c1**	**S2**	**S1**	**S1c1**	**S2c1**
LibriSpeech	female	0.492	0.087	0.092	0.072	0.267	0.246	0.237	0.614	0.115	0.121	0.305	0.121	0.318	0.069	0.067	0.052	0.053	0.139	0.053
	male	0.696	0.100	0.151	0.091	0.452	0.444	0.310	0.654	0.138	0.146	0.341	0.119	0.346	0.036	0.038	0.021	0.045	0.109	0.025
VCTK different	female	0.646	0.156	0.157	0.164	0.349	0.385	0.346	0.595	0.209	0.191	0.440	0.200	0.440	0.110	0.119	0.113	0.198	0.351	0.116
	male	0.682	0.111	0.180	0.110	0.433	0.398	0.436	0.492	0.096	0.084	0.461	0.121	0.460	0.042	0.043	0.069	0.076	0.132	0.081
VCTK common	female	0.653	0.210	0.203	0.179	0.447	0.457	0.344	0.614	0.089	0.098	0.446	0.199	0.450	0.107	0.115	0.127	0.154	0.252	0.136
	male	0.683	0.127	0.207	0.110	0.487	0.429	0.306	0.579	0.048	0.054	0.483	0.124	0.482	0.029	0.034	0.043	0.047	0.148	0.053

ZEBRA: worst-case privacy disclosure (individual) – Ignorant (oa)																				
Data	**Gender**	**Orig**	**A1**	**A2**	**B1**	**B2**	**D1**	**I1**	**K2**	**M1**	**M1c1**	**M1c2**	**M1c3**	**M1c4**	**O1**	**O1c1**	**S2**	**S1**	**S1c1**	**S2c1**
LibriSpeech	female	3.829	0.629	0.344	0.310	3.129	3.219	2.161	1.301	0.556	0.948	2.082	0.435	2.112	0.737	1.617	1.316	0.976	1.455	1.015
	male	4.055	0.016	0.476	0.258	3.062	3.046	3.622	1.620	1.386	1.354	2.124	0.355	2.298	0.984	0.997	0.786	0.997	1.775	0.956
VCTK different	female	3.972	1.473	1.059	1.570	1.825	1.318	2.908	0.171	1.172	0.929	2.192	0.999	2.193	1.752	1.478	1.221	1.472	2.118	1.524
	male	4.037	0.332	0.000	0.000	1.888	1.818	3.312	1.444	0.411	0.291	2.064	0.270	2.020	0.134	0.383	0.145	0.168	1.110	0.188
VCTK common	female	3.596	0.845	0.322	0.229	1.447	1.748	2.100	0.628	0.294	0.434	1.808	0.515	1.991	1.117	1.212	0.763	0.706	1.201	0.837
	male	3.616	1.322	0.067	0.368	1.924	2.447	2.779	0.434	1.146	0.602	2.608	0.669	2.488	0.535	0.845	1.146	0.477	1.447	1.447
ZEBRA: worst-case privacy disclosure (individual) – Lazy-informed (aa)																				
Data	**Gender**	**Orig**	**A1**	**A2**	**B1**	**B2**	**D1**	**I1**	**K2**	**M1**	**M1c1**	**M1c2**	**M1c3**	**M1c4**	**O1**	**O1c1**	**S2**	**S1**	**S1c1**	**S2c1**
LibriSpeech	female	3.829	1.617	1.617	2.094	3.531	3.520	3.208	2.815	2.094	2.316	2.677	2.069	3.023	1.839	2.094	1.617	1.219	2.520	1.793
	male	4.055	1.900	1.738	1.951	3.673	3.684	3.415	3.190	1.979	2.502	3.014	1.997	2.889	0.934	0.868	0.821	1.298	2.076	0.997
VCTK different	female	3.972	1.180	1.800	1.473	3.232	3.389	3.088	3.359	2.504	2.149	2.881	1.774	2.998	1.328	1.707	1.237	2.141	2.884	1.415
	male	4.037	1.446	1.851	1.462	3.549	3.316	3.773	3.156	1.411	1.110	2.626	1.779	2.623	1.110	1.110	1.508	1.622	1.809	1.353
VCTK common	female	3.596	1.498	1.690	1.514	2.845	2.779	2.376	3.298	1.447	1.447	3.100	1.924	2.972	1.845	1.447	1.393	1.447	2.100	1.499
	male	3.616	1.447	1.753	1.748	2.862	2.508	2.997	2.479	1.447	1.623	2.294	1.690	2.322	1.146	1.146	0.845	1.322	1.773	1.146

Table 5: Objective results: ZEBRA expected and worst-case privacy disclosure for ignorant and lazy-informed attack models on the test data.

ZEBRA: Expected privacy disclosure (population), bit – Ignorant (oa)																				
Data	**Gender**	**Orig**	**A1**	**A2**	**B1**	**B2**	**D1**	**I1**	**K2**	**M1**	**M1c1**	**M1c2**	**M1c3**	**M1c4**	**O1**	**O1c1**	**S2**	**S1**	**S1c1**	**S2c1**
LibriSpeech	female	0.584	0.004	0.003	0.004	0.221	0.215	0.268	0.044	0.003	0.002	0.213	0.009	0.257	0.021	0.018	0.014	0.017	0.031	0.019
	male	0.690	0.002	0.000	0.001	0.354	0.359	0.353	0.072	0.002	0.001	0.272	0.001	0.277	0.011	0.012	0.010	0.012	0.045	0.013
VCTK different	female	0.594	0.000	0.000	0.001	0.141	0.138	0.178	0.000	0.000	0.000	0.195	0.005	0.179	0.013	0.015	0.003	0.005	0.011	0.002
	male	0.667	0.000	0.000	0.000	0.196	0.190	0.101	0.007	0.000	0.000	0.217	0.000	0.222	0.006	0.006	0.002	0.003	0.012	0.003
VCTK common	female	0.653	0.007	0.006	0.004	0.132	0.138	0.161	0.003	0.001	0.001	0.185	0.007	0.181	0.016	0.015	0.013	0.017	0.037	0.012
	male	0.694	0.001	0.000	0.000	0.199	0.196	0.212	0.019	0.001	0.001	0.341	0.000	0.346	0.008	0.009	0.009	0.007	0.040	0.010
ZEBRA: Expected privacy disclosure (population), bit – Lazy-informed (aa)																				
Data	**Gender**	**Orig**	**A1**	**A2**	**B1**	**B2**	**D1**	**I1**	**K2**	**M1**	**M1c1**	**M1c2**	**M1c3**	**M1c4**	**O1**	**O1c1**	**S2**	**S1**	**S1c1**	**S2c1**
LibriSpeech	female	0.584	0.162	0.153	0.111	0.356	0.341	0.286	0.702	0.110	0.128	0.250	0.107	0.327	0.058	0.049	0.071	0.068	0.146	0.078
	male	0.690	0.084	0.133	0.066	0.524	0.514	0.401	0.617	0.110	0.123	0.335	0.070	0.315	0.027	0.023	0.015	0.042	0.061	0.015
VCTK different	female	0.594	0.107	0.087	0.104	0.315	0.293	0.170	0.635	0.129	0.120	0.390	0.135	0.359	0.094	0.096	0.071	0.111	0.150	0.080
	male	0.667	0.103	0.230	0.113	0.424	0.404	0.231	0.548	0.093	0.093	0.424	0.108	0.399	0.021	0.023	0.031	0.038	0.114	0.035
VCTK common	female	0.653	0.149	0.136	0.117	0.377	0.352	0.210	0.648	0.077	0.083	0.397	0.155	0.383	0.062	0.067	0.055	0.085	0.142	0.059
	male	0.694	0.159	0.268	0.113	0.458	0.434	0.384	0.568	0.065	0.079	0.558	0.126	0.538	0.029	0.027	0.039	0.053	0.137	0.049

ZEBRA: worst-case privacy disclosure (individual) – Ignorant (oa)																				
Data	**Gender**	**Orig**	**A1**	**A2**	**B1**	**B2**	**D1**	**I1**	**K2**	**M1**	**M1c1**	**M1c2**	**M1c3**	**M1c4**	**O1**	**O1c1**	**S2**	**S1**	**S1c1**	**S2c1**
LibriSpeech	female	3.979	0.486	0.496	0.310	2.266	2.191	3.035	1.476	0.766	0.657	2.288	0.962	2.265	0.389	0.399	1.310	0.823	0.804	0.708
	male	3.924	0.202	0.001	0.282	2.614	2.656	2.947	1.866	0.956	0.721	1.966	0.302	1.908	0.691	1.022	0.545	0.710	1.217	0.662
VCTK different	female	3.655	0.371	0.651	0.128	2.395	1.772	1.617	0.000	0.005	0.319	2.304	0.127	2.274	0.526	0.371	0.625	0.485	0.854	0.827
	male	3.921	0.000	0.000	0.000	3.012	3.063	2.160	1.043	0.119	0.073	2.413	0.000	2.344	1.182	1.182	0.663	0.140	0.566	0.663
VCTK common	female	3.557	0.423	0.741	0.668	1.197	1.153	2.187	0.386	0.102	0.095	2.187	0.367	2.100	0.470	0.559	1.447	0.706	1.117	1.146
	male	3.675	0.447	0.192	0.447	2.488	2.690	2.909	1.204	0.243	0.183	2.468	0.544	2.401	0.544	1.146	0.380	0.669	0.720	0.392
ZEBRA: worst-case privacy disclosure (individual) – Lazy-informed (aa)																				
Data	**Gender**	**Orig**	**A1**	**A2**	**B1**	**B2**	**D1**	**I1**	**K2**	**M1**	**M1c1**	**M1c2**	**M1c3**	**M1c4**	**O1**	**O1c1**	**S2**	**S1**	**S1c1**	**S2c1**
LibriSpeech	female	3.979	2.310	2.389	2.486	3.580	3.453	3.174	4.035	1.854	2.051	1.912	1.131	3.143	2.088	1.912	2.213	2.310	2.802	2.265
	male	3.924	2.169	1.801	2.403	3.525	3.609	3.031	2.631	1.868	2.022	2.936	2.102	2.829	0.841	0.546	1.022	1.659	1.721	0.926
VCTK different	female	3.655	1.429	1.869	1.869	2.598	3.020	1.781	3.819	1.429	1.827	2.290	1.656	2.204	1.526	1.781	1.304	1.429	1.702	1.128
	male	3.921	1.124	1.756	1.240	3.486	3.461	2.056	2.964	1.483	1.359	2.619	1.315	2.359	0.881	0.685	0.564	1.359	1.487	1.182
VCTK common	female	3.557	1.447	1.748	1.447	2.216	2.488	2.187	3.141	1.447	1.146	2.267	1.748	2.270	1.172	1.208	1.447	0.845	1.229	1.447
	male	3.675	1.447	1.857	1.447	3.080	2.702	2.157	2.593	1.021	1.447	2.902	1.170	2.561	0.869	1.146	0.502	1.146	1.447	0.618

3.3. Linkability

The *linkability* metric was proposed by Gomez-Barrero et al. (2017) for biometric template protection systems[4] and has been recently applied for the speech anonymization task by Maouche et al. (2020).

According to Gomez-Barrero et al. (2017), the local measure denoted $D_{\leftrightarrow}(s) \in [0, 1]$ — a system score-wise linkability, evaluates the linkability of a system for a given specific linkage score. The local linkability metric for score s is defined as

$$D_{\leftrightarrow}(s) = \max\{0, p(H|s) - p(\bar{H}|s)\}, \tag{2}$$

where variables H and $\bar{H}$ express whether two random utterances belong to the same speaker (*target*) or to different speakers (*impostor*) respectively.

The global linkability metric $D_{\leftrightarrow}^{\text{sys}}$ is calculated over all target scores:

$$D_{\leftrightarrow}^{\text{sys}} = \int p(s|H) \cdot D_{\leftrightarrow}(s)\, ds. \tag{3}$$

The global linkability metric $D_{\leftrightarrow}^{\text{sys}} \in [0, 1]$ provides an estimation of the global linkability of a system across all scores. It evaluates how non-overlapping are the score distributions of target and impostor pairs. A

[4]Definition of linkability (Gomez-Barrero et al., 2017): "*two templates are fully linkable if there exists some method to decide that they were extracted, with all certainty, from the same biometric instance. Two templates are linkable to a certain degree if there exists some method to decide that it is more likely that they were extracted from the same instance than from different instances.*"

linkability value of 0 means that the two distributions are indistinguishable, hence scores cannot be exploited by attackers and perfect privacy is achieved.

Results. Tables 6 and 7 present the linkability values for ignorant and lazy-informed attackers on the development and test datasets. Linkability plots with target/impostor score distributions for the three attack models are shown in Figure 7.

Table 6: Objective results: Linkability for ignorant and lazy-informed attack models on the development data.

Linkability – Ignorant (oa)																				
Data	**Gender**	**Orig**	**A1**	**A2**	**B1**	**B2**	**D1**	**I1**	**K2**	**M1**	**M1c1**	**M1c2**	**M1c3**	**M1c4**	**O1**	**O1c1**	**S2**	**S1**	**S1c1**	**S2c1**
LibriSpeech	female	0.801	0.076	0.067	0.088	0.258	0.299	0.326	0.132	0.085	0.085	0.376	0.052	0.389	0.105	0.108	0.089	0.111	0.099	0.093
	male	0.974	0.082	0.128	0.105	0.563	0.519	0.681	0.156	0.107	0.117	0.420	0.096	0.431	0.074	0.092	0.063	0.085	0.141	0.066
VCTK different	female	0.931	0.102	0.064	0.059	0.215	0.237	0.667	0.084	0.074	0.078	0.371	0.061	0.393	0.114	0.108	0.120	0.073	0.158	0.118
	male	0.968	0.058	0.081	0.069	0.354	0.375	0.476	0.092	0.053	0.055	0.566	0.068	0.565	0.075	0.072	0.069	0.044	0.138	0.083
VCTK common	female	0.935	0.055	0.048	0.084	0.246	0.262	0.593	0.142	0.049	0.065	0.311	0.063	0.304	0.093	0.094	0.082	0.073	0.116	0.089
	male	0.963	0.094	0.114	0.094	0.434	0.416	0.464	0.106	0.074	0.062	0.513	0.074	0.498	0.067	0.054	0.063	0.056	0.111	0.077
Linkability – Lazy-informed (aa)																				
Data	**Gender**	**Orig**	**A1**	**A2**	**B1**	**B2**	**D1**	**I1**	**K2**	**M1**	**M1c1**	**M1c2**	**M1c3**	**M1c4**	**O1**	**O1c1**	**S2**	**S1**	**S1c1**	**S2c1**
LibriSpeech	female	0.801	0.260	0.268	0.223	0.498	0.485	0.438	0.894	0.272	0.296	0.535	0.262	0.553	0.219	0.206	0.160	0.230	0.304	0.161
	male	0.974	0.261	0.331	0.245	0.757	0.748	0.577	0.932	0.350	0.360	0.593	0.284	0.599	0.121	0.124	0.095	0.150	0.225	0.107
VCTK different	female	0.931	0.368	0.378	0.387	0.634	0.667	0.617	0.886	0.459	0.430	0.733	0.442	0.730	0.281	0.289	0.285	0.408	0.576	0.292
	male	0.968	0.306	0.402	0.300	0.745	0.683	0.703	0.766	0.282	0.270	0.748	0.318	0.744	0.152	0.156	0.229	0.236	0.328	0.251
VCTK common	female	0.935	0.415	0.414	0.364	0.731	0.732	0.582	0.885	0.260	0.271	0.711	0.402	0.708	0.256	0.275	0.294	0.331	0.438	0.302
	male	0.963	0.279	0.370	0.245	0.746	0.686	0.595	0.852	0.168	0.178	0.754	0.274	0.744	0.125	0.116	0.150	0.140	0.290	0.171

Table 7: Objective results: Linkability for ignorant and lazy-informed attack models on the test data.

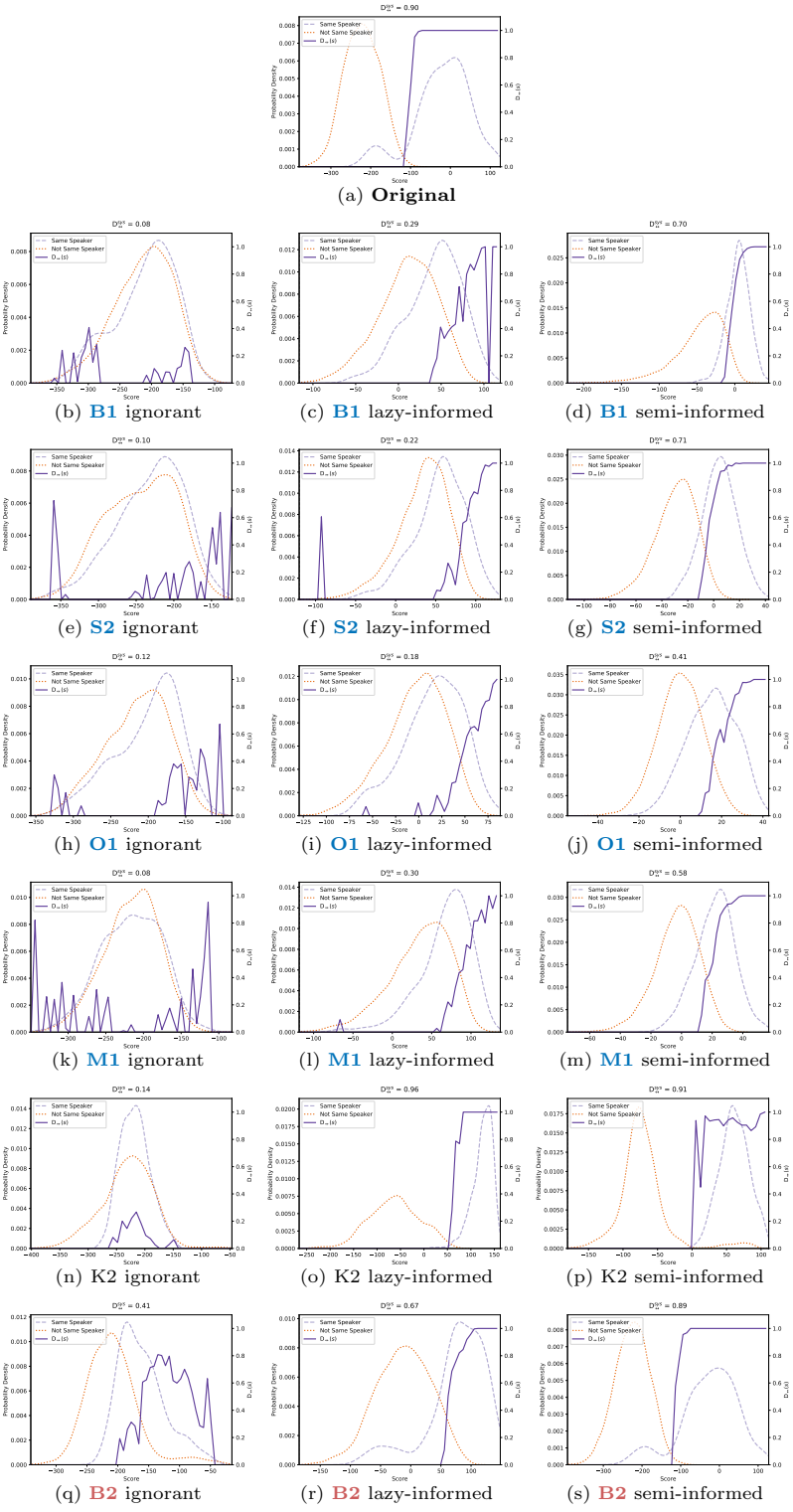
Linkability – Ignorant (oa)																				
Data	**Gender**	**Orig**	**A1**	**A2**	**B1**	**B2**	**D1**	**I1**	**K2**	**M1**	**M1c1**	**M1c2**	**M1c3**	**M1c4**	**O1**	**O1c1**	**S2**	**S1**	**S1c1**	**S2c1**
LibriSpeech	female	0.898	0.067	0.075	0.076	0.412	0.402	0.496	0.163	0.081	0.078	0.385	0.090	0.479	0.122	0.122	0.101	0.099	0.133	0.096
	male	0.958	0.074	0.085	0.083	0.578	0.579	0.654	0.142	0.083	0.089	0.527	0.082	0.521	0.085	0.071	0.084	0.101	0.141	0.086
VCTK different	female	0.881	0.045	0.050	0.056	0.379	0.359	0.370	0.172	0.062	0.060	0.393	0.069	0.375	0.091	0.091	0.042	0.060	0.077	0.046
	male	0.950	0.055	0.073	0.059	0.375	0.371	0.270	0.191	0.087	0.082	0.454	0.064	0.461	0.062	0.059	0.049	0.053	0.084	0.045
VCTK common	female	0.924	0.084	0.062	0.075	0.301	0.305	0.322	0.097	0.054	0.076	0.363	0.081	0.367	0.081	0.092	0.066	0.078	0.118	0.063
	male	0.972	0.079	0.099	0.074	0.420	0.388	0.417	0.173	0.072	0.070	0.584	0.082	0.577	0.057	0.078	0.069	0.072	0.159	0.082
Linkability – Lazy-informed (aa)																				
Data	**Gender**	**Orig**	**A1**	**A2**	**B1**	**B2**	**D1**	**I1**	**K2**	**M1**	**M1c1**	**M1c2**	**M1c3**	**M1c4**	**O1**	**O1c1**	**S2**	**S1**	**S1c1**	**S2c1**
LibriSpeech	female	0.898	0.346	0.355	0.295	0.666	0.654	0.587	0.962	0.298	0.320	0.542	0.319	0.603	0.183	0.164	0.223	0.278	0.305	0.238
	male	0.958	0.217	0.281	0.192	0.800	0.779	0.625	0.873	0.274	0.293	0.616	0.200	0.591	0.124	0.106	0.073	0.134	0.190	0.090
VCTK different	female	0.881	0.295	0.275	0.281	0.604	0.571	0.394	0.920	0.347	0.331	0.675	0.323	0.638	0.253	0.253	0.220	0.296	0.351	0.235
	male	0.950	0.282	0.463	0.295	0.729	0.712	0.458	0.845	0.273	0.280	0.704	0.285	0.673	0.104	0.109	0.138	0.149	0.287	0.146
VCTK common	female	0.924	0.318	0.324	0.275	0.636	0.588	0.432	0.919	0.224	0.245	0.618	0.329	0.595	0.160	0.169	0.167	0.243	0.321	0.173
	male	0.972	0.349	0.495	0.276	0.723	0.689	0.638	0.839	0.197	0.233	0.820	0.304	0.800	0.097	0.111	0.151	0.172	0.330	0.182

3.4. Voice similarity matrices

To visualize anonymization performance across different speakers in a dataset, *voice similarity matrices* have been proposed by Noé et al. (2020). A voice similarity matrix $M = (S(i,j))_{1\leq i\leq N, 1\leq j\leq N}$ is defined for a set of N speakers using similarity values $S(i,j)$ computed for speakers i and j as follows:

$$S(i,j) = \text{sigmoid}\left(\frac{1}{n_i n_j}\sum_{\substack{1\leq k\leq n_i \text{ and } 1\leq l\leq n_j \\ k\neq l \text{ if } n_j=n_j}} LLR(x_k^{(i)}, x_l^{(j)})\right), \tag{4}$$

where $LLR(x_k^{(i)}, x_l^{(j)})$ is the log-likelihood-ratio from the comparison of the k-th segment from the i-th speaker and the l-th segment from the j-th speaker, n_i, n_j are the numbers of segments for the corresponding speakers. Three types of matrices are computed: M_{oo} – on original data; M_{aa} – on anonymized data; and



(a) **Original**

(b) **B1** ignorant (c) **B1** lazy-informed (d) **B1** semi-informed

(e) **S2** ignorant (f) **S2** lazy-informed (g) **S2** semi-informed

(h) **O1** ignorant (i) **O1** lazy-informed (j) **O1** semi-informed

(k) **M1** ignorant (l) **M1** lazy-informed (m) **M1** semi-informed

(n) K2 ignorant (o) K2 lazy-informed (p) K2 semi-informed

(q) **B2** ignorant (r) **B2** lazy-informed (s) **B2** semi-informed

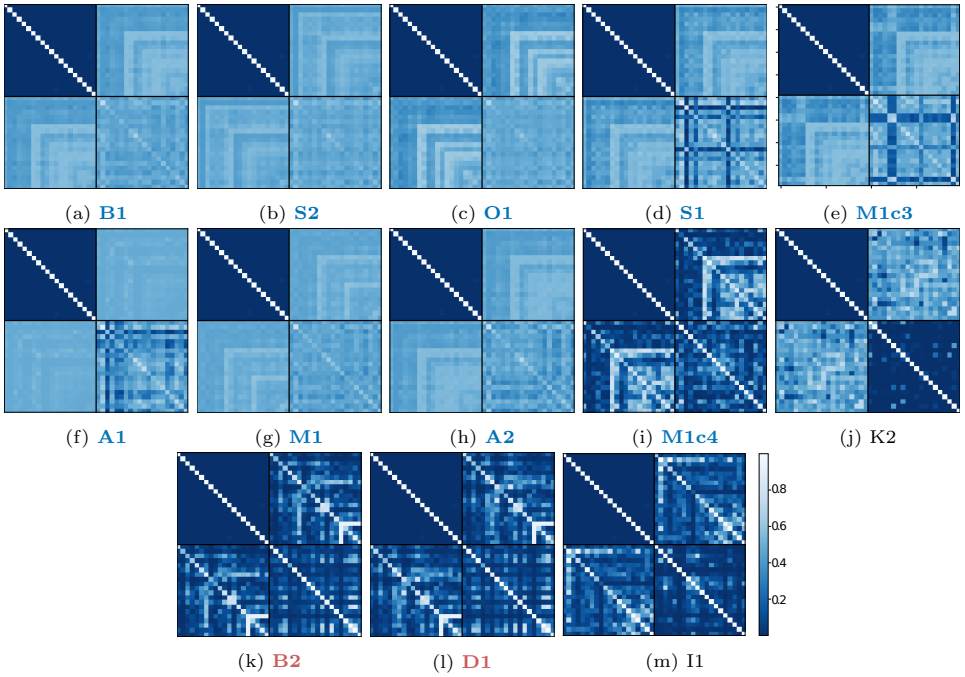
Figure 7: Female speaker linkability results on *LibriSpeech-test* computed for the selected set of primary anonymization systems.

M_{oa} – on original and anonymized data. In the latter case, for computing $S(i, j)$, we use original data for speaker i and anonymized data for speaker j. A global matrix M is then constructed as

$$M = \begin{pmatrix} M_{oo} & M_{oa} \\ M_{oa} & M_{aa} \end{pmatrix}. \tag{5}$$

The voice similarity matrices shown in Figures 8, 9 for *LibriSpeech-test* and in Figures 10, 11 for *VCTK-test (different)* show substantial differences between the submitted systems. For M_{oo}, a distinct diagonal in the similarity matrix points out the speaker discrimination ability in the original set, while in M_{oa}, the diagonal disappears if the protection is good. In M_{aa}, the diagonal of the matrix emerges if the resulting pseudo-voices can be distinguished (Noé et al., 2020). The matrices for signal-processing based approaches (**B2,D1,I**) exhibit a distinct diagonal for M_{aa} matrices, indicating that voices remain distinguishable after anonymization. Among x-vector based systems, a distinct diagonal for M_{aa} is observed only for system **K2**. For system **M1c4**, where x-vectors provided to SS AM are original, a distinct diagonal is observed for M_{aa} and a less distinct one for M_{oa}. For x-vector based anonymization systems related to **B1**, no diagonal is observed for M_{oa} for all datasets which suggests high de-identification performance. System **K2** has the most distinct M_{aa} diagonal on *LibriSpeech-test* (Figures 8j and 9j), and confusions between some speaker voices can be seen only for male speakers (Figure 9j). For *VCTK-test (differnt)*, both matrices M_{oo} and M_{aa} have less distinct diagonals in comparison to *LibriSpeech-test*.

All the confusion matrices considered above were computed using the ASV_{eval} model trained on original (non-anonymized data). If we retrain the speaker verification model on the anonymized data and re-compute the confusion matrices using the obtained $ASR_{\text{eval}}^{\text{anon}}$ model, we can observe in Figure 12 that the diagonals of the M_{aa} matrices become much more distinct. This means that the voice distinctiveness is better preserved with respect to the $ASR_{\text{eval}}^{\text{anon}}$ model than with respect to ASR_{eval}.

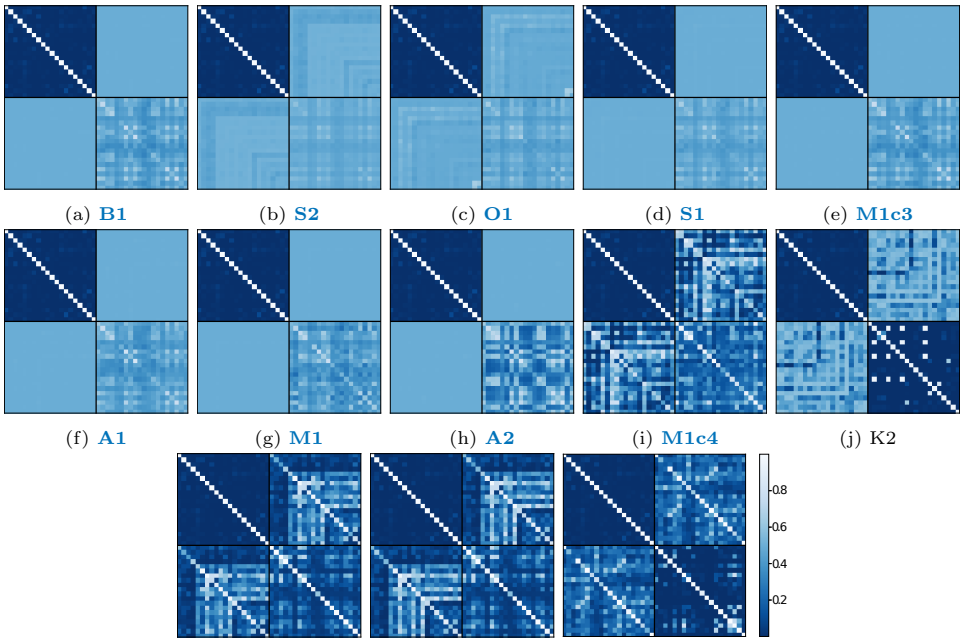


(a) **B1** (b) **S2** (c) **O1** (d) **S1** (e) **M1c3**

(f) **A1** (g) **M1** (h) **A2** (i) **M1c4** (j) K2

(k) **B2** (l) **D1** (m) I1

Figure 8: Voice similarity matrices on *LibriSpeech-test* for female speakers.

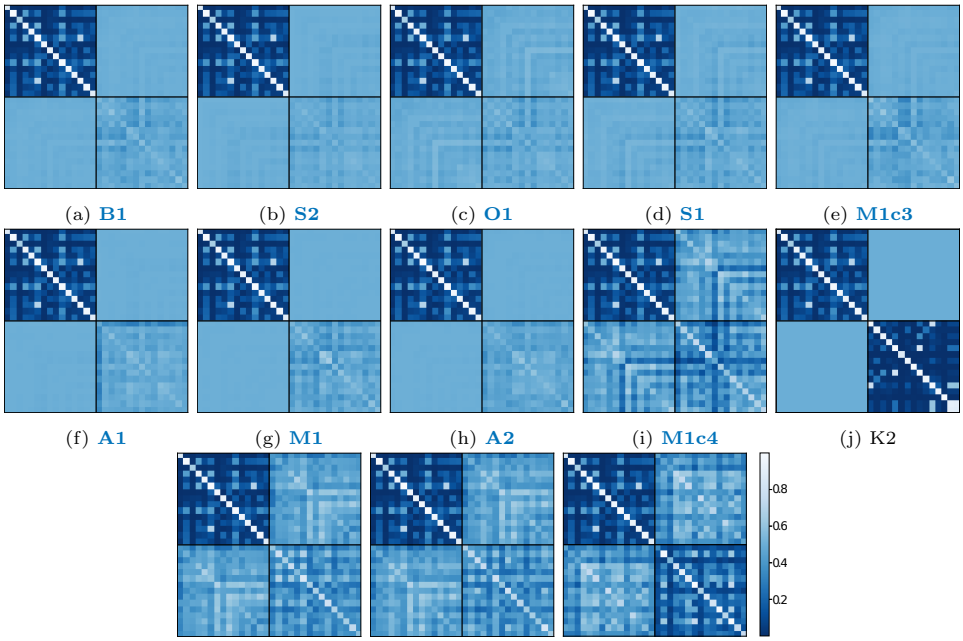


(a) **B1** (b) **S2** (c) **O1** (d) **S1** (e) **M1c3**

(f) **A1** (g) **M1** (h) **A2** (i) **M1c4** (j) K2

(k) **B2** (l) **D1** (m) I1

Figure 9: Voice similarity matrices on *LibriSpeech-test* for male speakers.

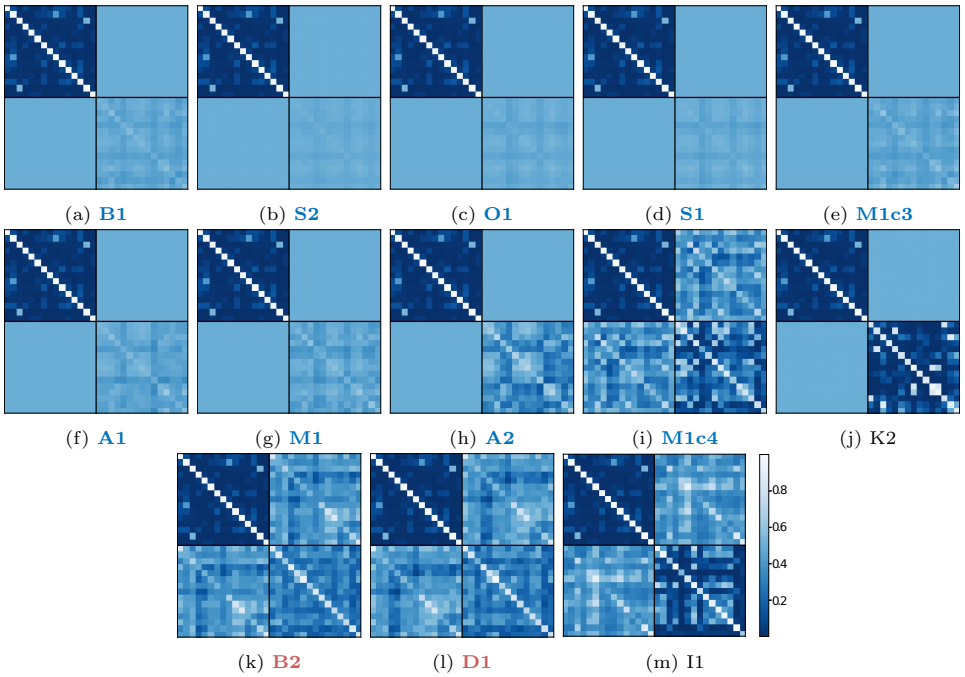


(a) **B1** (b) **S2** (c) **O1** (d) **S1** (e) **M1c3**

(f) **A1** (g) **M1** (h) **A2** (i) **M1c4** (j) K2

(k) **B2** (l) **D1** (m) I1

Figure 10: Voice similarity matrices on *VCTK-test (different)* for female speakers.

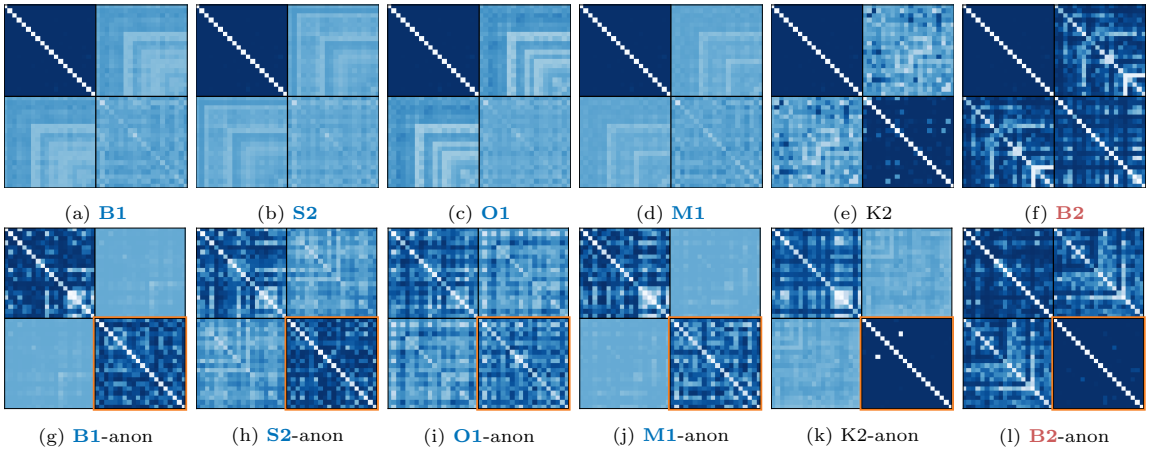


(a) **B1** (b) **S2** (c) **O1** (d) **S1** (e) **M1c3**

(f) **A1** (g) **M1** (h) **A2** (i) **M1c4** (j) K2

(k) **B2** (l) **D1** (m) I1

Figure 11: Voice similarity matrices on *VCTK-test (different)* for male speakers.



(a) **B1** (b) **S2** (c) **O1** (d) **M1** (e) K2 (f) **B2**

(g) **B1**-anon (h) **S2**-anon (i) **O1**-anon (j) **M1**-anon (k) K2-anon (l) **B2**-anon

Figure 12: Comparison of voice similarity matrices computed with (1) ASV_{eval} trained on original data (upper row) and (2) $\text{ASV}_{\text{eval}}^{\text{anon}}$ trained on anonymized speech data (lower row) on *LibriSpeech-test* computed for female speakers.

3.5. Gain of voice distinctiveness and de-identification

De-identification (DeID) and gain of voice distinctiveness (G_{VD}) metrics are computed from voice similarity matrices (Noé et al., 2020). They are estimated based on the ratio of diagonal dominance for a pair of matrices $\{M_{oa}, M_{oo}\}$ and $\{M_{oo}, M_{oo}\}$ correspondingly. The diagonal dominance $D_{\text{diag}}(M)$ is defined as

the absolute difference between the mean values of diagonal and off-diagonal elements:[5]

$$D_{\text{diag}}(M)=\left| \sum_{1\le i\le N} \frac{S(i,i)}{N} - \sum_{1\le j\le N \text{ and } j\neq k} \sum_{1\le k\le N} \frac{S(j,k)}{N(N-1)} \right|. \tag{6}$$

Gain of voice distinctiveness. Gain of voice distinctiveness is defined as:

$$G_{\text{VD}} = 10\log_{10} \frac{D_{\text{diag}}(M_{aa})}{D_{\text{diag}}(M_{oo})}, \tag{7}$$

where 0 means that the voice distinctiveness remains globally the same in the protected space, and gain above or below 0 corresponds respectively to increase or loss of global voice distinctiveness.

De-identification. De-identification is calculated as:

$$\text{DeID} = 1 - \frac{D_{\text{diag}}(M_{oa})}{D_{\text{diag}}(M_{oo})}. \tag{8}$$

DeID = 100% assumes perfect de-identification, while DeID = 0 corresponds to a system which achieves no de-identification.

Results. *Gain of voice distinctiveness (G_{VD})* results are presented in Figure 13 and Tables 8, 9. Anonymization leads to the loss of voice distinctiveness for all (except **K2**) anonymization systems. Signal-processing based methods much better preserve voice distinctiveness than methods related to **B1** and the best results are achieved for methods **K2** and **I1**. There is a gap in performance between male and female speakers for most of the systems: for some anonymization methods, G_{VD} is higher for female speakers than for male (in particular, for **S2, S2c1**, and others), while for other methods (i.e. **A2**), on the contrary, G_{VD} is lower for female than for male speakers.

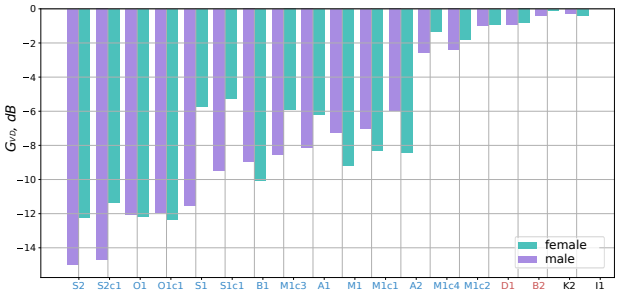
Results for the *de-identification* (DeId) metric are shown in Figure 14 and Tables 8, 9. For most of the x-vector based anonymization methods almost full de-identification is achieved. In comparison with x-vector based anonymization methods, all signal-processing based techniques demonstrate a considerably lower level of de-identification.

Table 8: Objective results: Gain of voice distinctiveness (G_{VD}) and De-identification (DeID) on the development data.

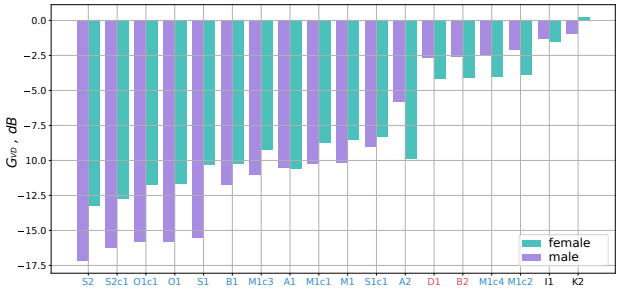
Gain of voice distinctiveness (G_{VD})																			
Data	**Gender**	**A1**	**A2**	**B1**	**B2**	**D1**	**I1**	**K2**	**M1**	**M1c1**	**M1c2**	**M1c3**	**M1c4**	**O1**	**O1c1**	**S2**	**S1**	**S1c1**	**S2c1**
LibriSpeech	female	-8.48	-8.09	-9.17	-1.06	-1.15	-0.99	-0.29	-8.14	-7.69	-1.81	-6.85	-1.38	-12.03	-11.96	-12.64	-5.72	-5.01	-11.90
	male	-7.93	-6.02	-8.76	-1.19	-1.26	-0.33	-0.20	-7.58	-7.16	-2.07	-7.28	-2.18	-13.36	-13.22	-14.81	-11.62	-7.55	-14.13
VCTK different	female	-8.85	-8.48	-8.81	-3.61	-3.24	-1.06	-0.45	-7.58	-7.81	-2.96	-7.69	-3.14	-8.95	-8.91	-10.33	-6.64	-5.02	-9.82
	male	-13.36	-8.86	-12.66	-2.98	-3.31	-0.60	-0.72	-11.71	-11.70	-2.88	-12.14	-3.09	-15.91	-15.94	-17.02	-15.35	-9.67	-15.68
VCTK common	female	-6.86	-6.49	-7.55	-1.40	-1.24	-0.25	-0.24	-7.10	-6.80	-1.85	-6.25	-2.06	-7.58	-7.68	-8.86	-6.06	-4.57	-8.29
	male	-9.91	-7.22	-10.42	-0.81	-0.72	-0.20	-0.60	-10.19	-10.18	-1.41	-9.42	-1.41	-14.23	-14.85	-16.34	-13.10	-7.52	-14.91

De-identification (DeID)																			
Data	**Gender**	**A1**	**A2**	**B1**	**B2**	**D1**	**I1**	**K2**	**M1**	**M1c1**	**M1c2**	**M1c3**	**M1c4**	**O1**	**O1c1**	**S2**	**S1**	**S1c1**	**S2c1**
LibriSpeech	female	98.7	99.6	99.6	55.4	58.0	54.0	97.7	99.9	99.8	49.1	99.3	43.3	97.0	96.9	96.9	97.7	93.9	96.7
	male	100.0	100.0	100.0	41.2	47.0	32.6	99.0	100.0	100.0	53.1	100.0	52.6	99.2	99.3	99.4	100.0	92.7	99.3
VCTK different	female	98.8	99.5	99.6	93.3	92.3	62.6	97.8	99.9	99.9	81.8	99.3	83.1	98.0	98.1	98.7	98.6	96.1	98.1
	male	100.0	100.0	100.0	71.2	72.5	79.2	99.2	100.0	100.0	73.4	100.0	74.5	99.8	99.8	99.8	99.8	97.9	99.8
VCTK common	female	98.8	99.8	99.5	84.1	83.6	50.4	97.6	99.2	99.4	70.9	98.9	70.8	96.5	96.9	98.0	98.1	92.9	96.5
	male	99.9	100.0	99.9	43.9	42.6	62.2	93.4	99.1	99.2	60.1	99.9	62.0	99.4	99.3	97.5	99.2	92.3	97.3

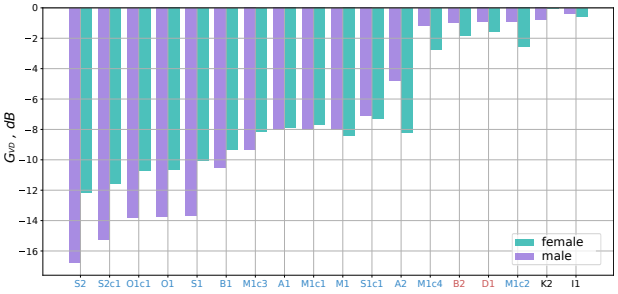
[5]See notations in Section 3.4



(a) LibriSpeech-test

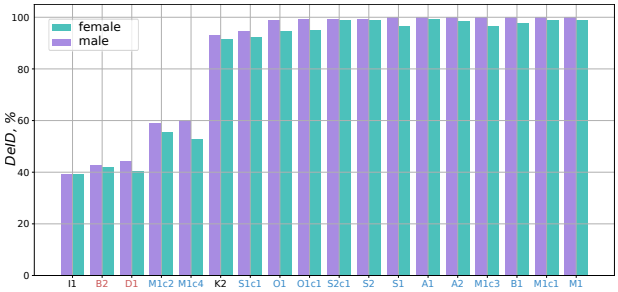


(b) VCTK-test (different)

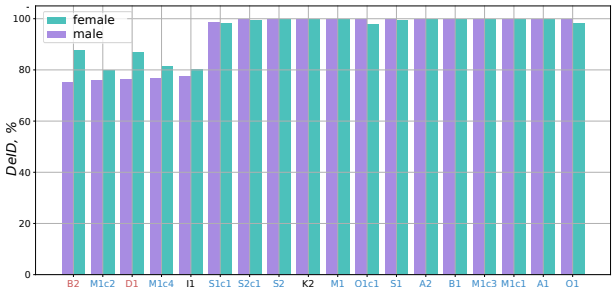


(c) VCTK-test (common)

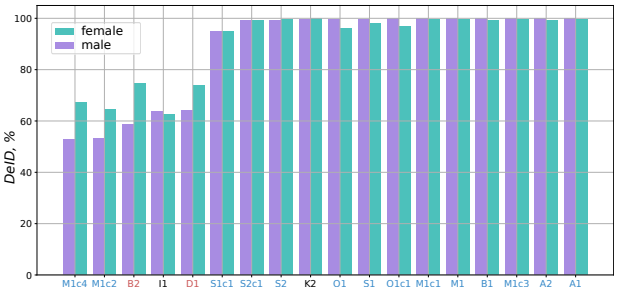
Figure 13: Gain of voice distinctiveness (G_{VD}) results on the test datasets for different anonymization systems. Blue and red colors in the system notations indicate systems developed from **B1** and **B2**, respectively. The results in each subfigure are ordered by metric values on male speakers. Higher G_{VD} values correspond to better voice distinctiveness preservation.



(a) LibriSpeech-test



(b) VCTK-test (different)



(c) VCTK-test (common)

Figure 14: De-identification (DeID) results on the test datasets for different anonymization systems. Blue and red colors in the system notations indicate systems developed from **B1** and **B2**, respectively. The results in each subfigure are ordered by metric values on male speakers. Higher DeID values correspond to better privacy.

Table 9: Objective results: Gain of voice distinctiveness (G_{VD}) and De-identification (DeID) on the test data.

Gain of voice distinctiveness (G_{VD})																			
Data	**Gender**	**A1**	**A2**	**B1**	**B2**	**D1**	**I1**	**K2**	**M1**	**M1c1**	**M1c2**	**M1c3**	**M1c4**	**O1**	**O1c1**	**S2**	**S1**	**S1c1**	**S2c1**
LibriSpeech	female	-6.21	-8.43	-10.07	-0.84	-0.93	-0.38	-0.10	-9.16	-8.32	-1.81	-5.93	-1.31	-12.17	-12.33	-12.21	-5.72	-5.27	-11.36
	male	-8.16	-5.98	-8.98	-0.90	-0.98	-0.26	-0.37	-7.25	-7.03	-2.42	-8.52	-2.55	-12.04	-11.95	-14.99	-11.54	-9.48	-14.71
VCTK different	female	-10.58	-9.87	-10.28	-4.11	-4.16	-1.51	0.25	-8.55	-8.78	-3.89	-9.26	-4.01	-11.67	-11.77	-13.25	-10.33	-8.35	-12.76
	male	-10.53	-5.83	-11.73	-2.61	-2.71	-1.29	-0.93	-10.21	-10.22	-2.10	-11.03	-2.47	-15.79	-15.81	-17.18	-15.56	-9.07	-16.24
VCTK common	female	-7.90	-8.20	-9.30	-1.80	-1.55	-0.56	-0.02	-8.43	-7.72	-2.57	-8.12	-2.75	-10.68	-10.69	-12.15	-10.03	-7.26	-11.56
	male	-8.04	-4.80	-10.49	-0.97	-0.92	-0.37	-0.80	-8.03	-8.03	-0.92	-9.35	-1.15	-13.77	-13.79	-16.76	-13.65	-7.12	-15.24

De-identification (DeID)																			
Data	**Gender**	**A1**	**A2**	**B1**	**B2**	**D1**	**I1**	**K2**	**M1**	**M1c1**	**M1c2**	**M1c3**	**M1c4**	**O1**	**O1c1**	**S2**	**S1**	**S1c1**	**S2c1**
LibriSpeech	female	99.5	98.5	97.9	41.9	40.5	39.3	91.5	99.0	99.1	55.8	96.6	52.8	94.7	95.0	98.8	96.5	92.5	99.0
	male	100.0	100.0	100.0	43.0	44.2	39.2	93.2	100.0	100.0	58.9	100.0	60.3	98.9	99.2	99.4	99.9	94.7	99.4
VCTK different	female	100.0	100.0	99.9	87.9	86.9	80.1	100.0	100.0	100.0	80.1	99.8	81.6	98.2	98.1	99.7	99.3	98.1	99.6
	male	100.0	100.0	100.0	75.0	76.5	77.7	100.0	100.0	100.0	76.2	100.0	76.7	100.0	100.0	100.0	100.0	98.7	100.0
VCTK common	female	99.6	99.3	99.3	74.6	73.7	62.6	99.9	99.7	99.7	64.5	99.5	67.3	96.0	96.9	99.4	98.2	94.8	99.3
	male	100.0	100.0	100.0	58.7	64.0	63.8	99.5	100.0	100.0	53.3	100.0	52.7	99.6	99.8	99.3	99.7	94.8	99.1

3.6. Using anonymized speech data for ASV training

Using anonymized speech data for ASV training leads to a stronger attack model referred to as *semi-informed* ($ASV_{\mathrm{eval}}^{\mathrm{anon}}$). The resulting privacy protection is assessed in Tables 10, 11 in terms of EER, C_{llr}, and $C_{\mathrm{llr}}^{\mathrm{min}}$, in Tables 12, 13 in terms of the two ZEBRA metrics, and in Tables 14, Table 15 in terms of linkability.

Table 10: Objective results: EER, C_{llr}, and $C_{\mathrm{llr}}^{\mathrm{min}}$ for the semi-informed attack model on the development data.

EER – Semi-informed (aa)								
Data	**Gender**	**Orig**	**B1**	**B2**	**O1**	**M1**	**S2**	**K2**
LibriSpeech	female	8.66	18.89	10.94	23.44	23.30	18.18	3.27
	male	1.24	7.45	1.09	14.60	24.22	11.18	1.55
VCTK different	female	2.86	12.41	3.54	14.94	14.94	13.03	1.91
	male	1.44	10.92	3.47	19.45	26.35	10.82	6.05
VCTK common	female	2.62	14.53	4.07	14.83	18.60	12.21	1.45
	male	1.43	16.81	3.13	22.51	32.19	13.68	3.13

C_{llr} – Semi-informed (aa)								
Data	**Gender**	**Orig**	**B1**	**B2**	**O1**	**M1**	**S2**	**K2**
LibriSpeech	female	42.9	6.9	46.2	2.6	4.0	4.4	1.5
	male	14.2	3.6	16.3	1.5	6.2	2.0	0.7
VCTK different	female	1.1	2.1	0.8	3.2	8.4	2.4	1.4
	male	1.2	2.2	1.1	3.0	16.0	3.3	2.2
VCTK common	female	0.9	1.6	1.4	1.9	6.6	1.2	0.5
	male	1.6	2.8	1.4	2.1	12.4	1.7	1.7

$C_{\mathrm{llr}}^{\mathrm{min}}$ – Semi-informed (aa)								
Data	**Gender**	**Orig**	**B1**	**B2**	**O1**	**M1**	**S2**	**K2**
LibriSpeech	female	0.304	0.563	0.351	0.666	0.627	0.566	0.114
	male	0.034	0.241	0.035	0.464	0.658	0.365	0.063
VCTK different	female	0.100	0.403	0.122	0.475	0.487	0.440	0.076
	male	0.052	0.373	0.127	0.602	0.764	0.376	0.223
VCTK common	female	0.088	0.473	0.143	0.484	0.532	0.373	0.043
	male	0.050	0.518	0.103	0.627	0.825	0.427	0.130

Table 11: Objective results: EER, C_{llr}, and $C_{\text{llr}}^{\text{min}}$ for the semi-informed attack model on the test data.

EER – Semi-informed (aa)								
Data	**Gender**	**Orig**	**B1**	**B2**	**O1**	**M1**	**S2**	**K2**
LibriSpeech	female	7.66	12.23	8.03	26.28	18.25	12.23	0.36
	male	1.11	10.69	1.56	16.48	23.39	11.14	3.79
VCTK different	female	4.89	16.20	9.05	20.37	22.22	17.70	1.59
	male	2.07	10.91	4.13	19.23	24.80	14.29	5.11
VCTK common	female	2.89	18.79	6.36	21.68	19.94	16.18	1.73
	male	1.13	13.28	2.54	16.67	27.12	13.28	3.11

C_{llr} – Semi-informed (aa)								
Data	**Gender**	**Orig**	**B1**	**B2**	**O1**	**M1**	**S2**	**K2**
LibriSpeech	female	26.8	3.0	30.4	3.7	3.3	2.5	0.1
	male	15.3	5.1	14.9	1.4	7.9	1.9	1.7
VCTK different	female	1.5	3.6	2.8	5.7	10.4	4.3	1.1
	male	1.8	2.2	1.9	5.4	15.3	4.2	2.7
VCTK common	female	0.9	2.0	1.4	3.6	7.8	2.3	0.6
	male	1.0	1.9	1.2	3.0	10.6	1.8	1.4

$C_{\text{llr}}^{\text{min}}$ – Semi-informed (aa)								
Data	**Gender**	**Orig**	**B1**	**B2**	**O1**	**M1**	**S2**	**K2**
LibriSpeech	female	0.183	0.384	0.204	0.726	0.548	0.384	0.011
	male	0.041	0.329	0.045	0.523	0.699	0.348	0.115
VCTK different	female	0.169	0.528	0.308	0.624	0.675	0.561	0.057
	male	0.072	0.368	0.147	0.597	0.722	0.453	0.199
VCTK common	female	0.091	0.552	0.211	0.614	0.608	0.496	0.066
	male	0.036	0.413	0.072	0.534	0.764	0.411	0.116

Table 12: Objective results: ZEBRA expected and worst-case privacy disclosure for the semi-informed attack model on the development data.

ZEBRA: Expected privacy disclosure (population) – Semi-informed								
Data	**Gender**	**Orig**	**B1**	**B2**	**O1**	**M1**	**S2**	**K2**
LibriSpeech	female	0.492	0.307	0.458	0.232	0.261	0.303	0.632
	male	0.696	0.541	0.694	0.376	0.239	0.449	0.671
VCTK different	female	0.646	0.421	0.630	0.368	0.359	0.392	0.663
	male	0.682	0.442	0.625	0.276	0.161	0.440	0.551
VCTK common	female	0.653	0.368	0.612	0.361	0.328	0.445	0.690
	male	0.683	0.337	0.644	0.261	0.120	0.404	0.620

ZEBRA: worst-case privacy disclosure (individual) – Semi-informed								
Data	**Gender**	**Orig**	**B1**	**B2**	**O1**	**M1**	**S2**	**K2**
LibriSpeech	female	3.829	2.884	3.801	2.884	2.949	3.015	2.829
	male	4.055	3.298	4.073	2.932	2.641	3.156	3.091
VCTK different	female	3.972	3.001	3.980	2.386	2.892	2.462	3.650
	male	4.037	2.632	3.910	2.110	1.985	2.850	3.208
VCTK common	female	3.596	2.690	3.513	2.401	2.271	2.286	3.660
	male	3.616	2.001	3.355	2.100	1.447	2.231	2.496

Table 13: Objective results: ZEBRA expected and worst-case privacy disclosure for the semi-informed attack model on the test data.

ZEBRA: Expected privacy disclosure (population) – Semi-informed								
Data	**Gender**	**Orig**	**B1**	**B2**	**O1**	**M1**	**S2**	**K2**
LibriSpeech	female	0.584	0.436	0.568	0.189	0.317	0.437	0.713
	male	0.690	0.477	0.687	0.333	0.207	0.461	0.630
VCTK different	female	0.594	0.329	0.491	0.260	0.223	0.305	0.678
	male	0.667	0.446	0.611	0.280	0.191	0.385	0.568
VCTK common	female	0.653	0.314	0.563	0.269	0.272	0.353	0.669
	male	0.694	0.414	0.668	0.324	0.163	0.416	0.631

ZEBRA: worst-case privacy disclosure (individual) – Semi-informed								
Data	**Gender**	**Orig**	**B1**	**B2**	**O1**	**M1**	**S2**	**K2**
LibriSpeech	female	3.979	3.235	3.966	2.690	3.074	3.155	3.967
	male	3.924	3.031	3.873	2.403	2.278	2.786	2.636
VCTK different	female	3.655	2.490	3.183	2.258	2.070	2.429	3.894
	female	3.921	2.589	3.844	2.296	1.743	3.137	3.113
VCTK common	female	3.557	2.350	2.924	1.718	1.748	1.804	2.748
	male	3.675	2.206	3.630	1.822	1.578	2.292	2.931

Table 14: Objective results: Linkability for the semi-informed attack model on the development data.

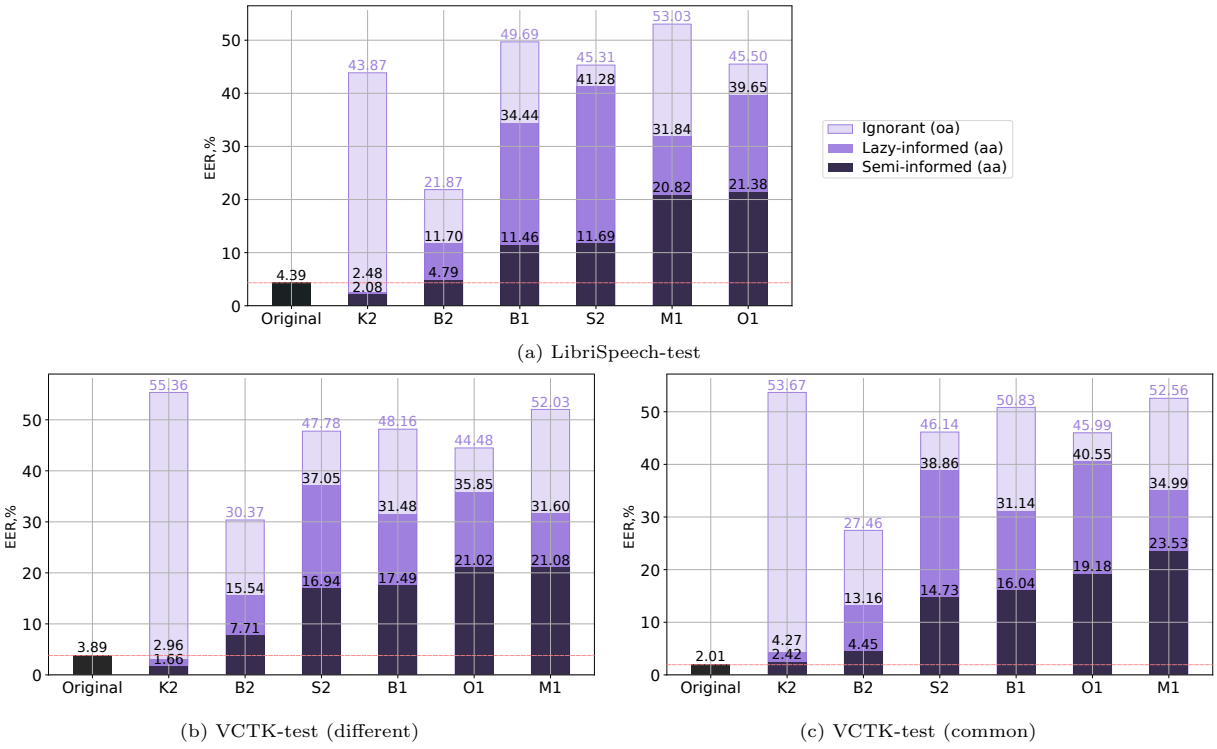
Linkability – Semi-informed								
Data	**Gender**	**Orig**	**B1**	**B2**	**O1**	**M1**	**S2**	**K2**
LibriSpeech	female	0.801	0.563	0.764	0.484	0.483	0.578	0.921
	male	0.974	0.814	0.977	0.652	0.455	0.725	0.956
VCTK different	female	0.931	0.700	0.915	0.645	0.650	0.677	0.952
	female	0.968	0.726	0.915	0.523	0.375	0.715	0.836
VCTK common	female	0.935	0.629	0.900	0.634	0.565	0.687	0.951
	male	0.963	0.589	0.914	0.460	0.289	0.644	0.910

Table 15: Objective results: Linkability for the semi-informed attack model on the test data.

Linkability – Semi-informed								
Data	**Gender**	**Orig**	**B1**	**B2**	**O1**	**M1**	**S2**	**K2**
LibriSpeech	female	0.898	0.700	0.891	0.408	0.577	0.710	0.988
	male	0.958	0.689	0.964	0.606	0.448	0.743	0.915
VCTK different	female	0.881	0.609	0.773	0.512	0.480	0.574	0.960
	female	0.950	0.727	0.901	0.531	0.407	0.650	0.865
VCTK common	female	0.924	0.540	0.834	0.475	0.491	0.602	0.955
	male	0.972	0.659	0.933	0.583	0.347	0.663	0.909

Figure 15 shows EER results for the semi-informed (darker, lower bars), lazy-informed, and ignorant attack models on (a) *LibriSpeech-test* and (b) *VCTK-test*. For two test sets (*LibriSpeech-test* and *VCTK-test (different)*), system **K2** even delivers lower EERs for the semi-informed attack model than for original data without anonymization. For the semi-informed attack model, x-vector based anonymization techniques related to **B1** (**O1, M1, S2**) demonstrate higher EER than other considered approaches (**B2, K2**). The best results against the semi-informed attack model are obtained by systems **M1** and **O1**.

Figure 16 shows mean EER results (over all VoicePrivacy development and test datasets) separately for male and female speakers. Speaker anonymization performs differently for male and female speakers. For system **M1**, anonymization for male speakers works better than for female speakers for all attack models, and for **B2**, on the contrary, results for female speakers are better.



(a) LibriSpeech-test

(b) VCTK-test (different)

(c) VCTK-test (common)

Figure 15: EER results achieved by different anonymization systems on the test datasets against the three attack models, compared to the EER achieved on the original data.

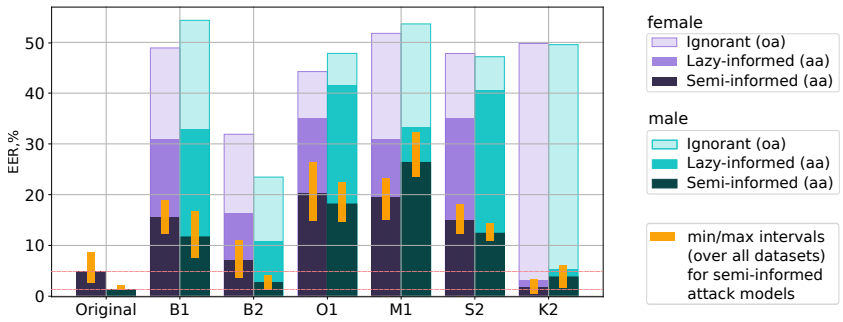


Figure 16: Mean EER results achieved by different anonymization systems over all development and test datasets for female and male speakers against the three attack models, compared to the EER achieved on the original data.

3.7. Comparison of privacy metrics

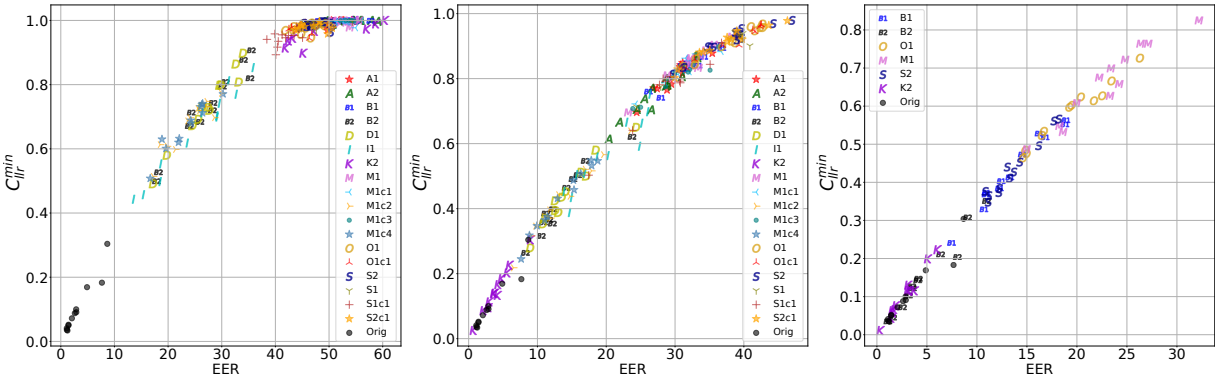
The considered privacy metrics correlate with each other to a variable extent. In this section, we investigate this observation in more detail. Figure 17 presents EER vs. $C_{\text{llr}}^{\text{min}}$ results for ignorant, lazy-informed, and semi-informed attack models for different datasets and anonymization systems, as well as for original (non-anonymized) data. Similarly, Figure 18 demonstrates the relation between linkability and $C_{\text{llr}}^{\text{min}}$ metrics. We observe a consistent correlation between all three metrics, especially for the lazy-informed and semi-informed attack models.

Figure 19 shows scatterplots for (EER, C_{llr}) for the three attack models. For the ignorant attack model (Figure 19a), the results for signal-processing methods (**B2,D1,I1**) and x-vector based methods form two separate clusters.

Figure 20 demonstrates the relation between linkability and EER. Both metrics perform similarly in most cases for lazy-informed and semi-informed attackers, though for the ignorant attacker they behave differently in some particular cases, e.g., for system **K2**. There are cases where the EER is above 50% that can be considered as perfect privacy, while linkability for these cases is higher than 0 meaning that according to linkability there is still some exploitable information for attackers left in the scores.

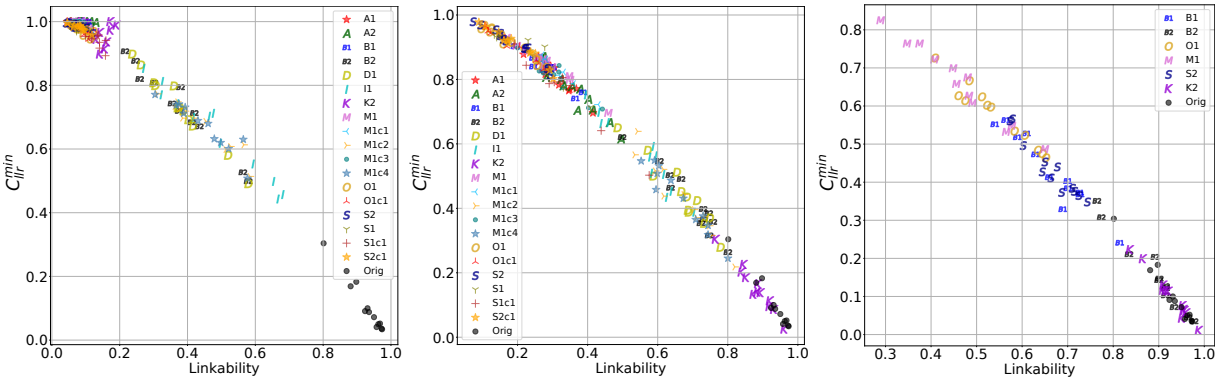
ZEBRA expected privacy disclosure (population) and $C_{\text{llr}}^{\text{min}}$ have a linear dependency as shown in Figure 21. ZEBRA worst case privacy disclosure (individual) differs from all the considered metrics as shown in Figure 22.

Finally, Figure 23 shows the relation between De-identification (DeID) on the one hand and Gain of voice distinctiveness (G_{VD}), EER, or $C_{\text{llr}}^{\text{min}}$ on the other hand. In particular, Figure 23a shows that methods derived from **B1** provide near-to-perfect de-identification, while the signal-processing anonymization solutions better preserve voice distinctiveness, and **K2** is the only system which reaches a good trade-off between de-identification performance and voice distinctiveness.



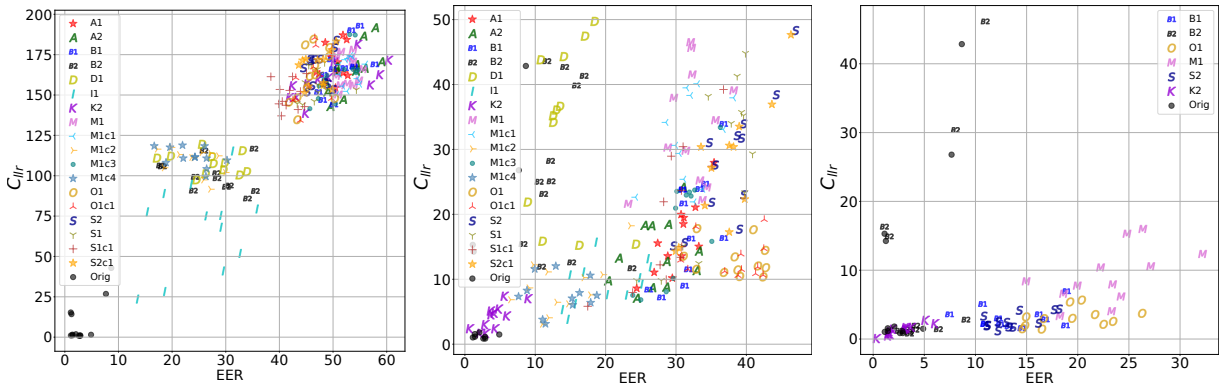
(a) Ignorant attacker (oa) (b) Lazy-informed attacker (aa) (c) Semi-informed attacker (aa)

Figure 17: EER vs. $C_{\text{llr}}^{\text{min}}$ results for the three attack models. Each point in the figure represents results on a dataset from the set of all 12 VoicePrivacy development and test datasets. Higher EER and $C_{\text{llr}}^{\text{min}}$ values correspond to better privacy.



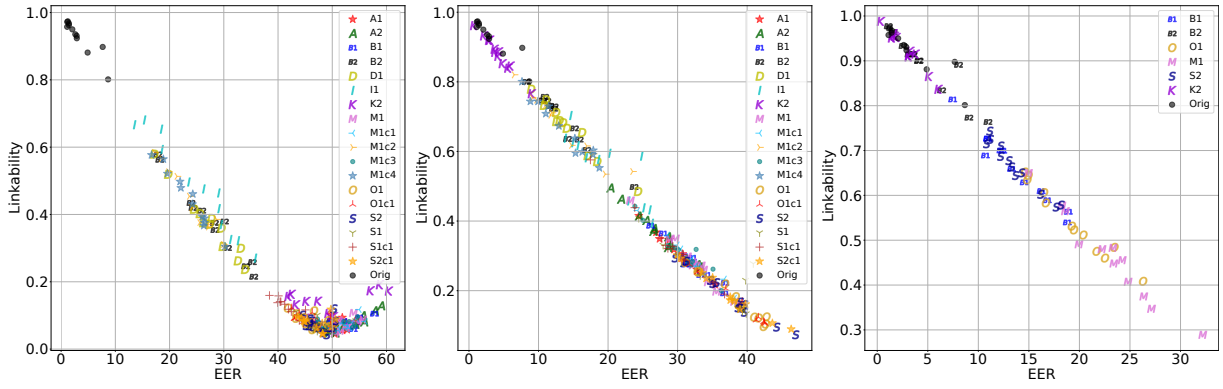
(a) Ignorant attacker (oa) (b) Lazy-informed attacker (aa) (c) Semi-informed attacker (aa)

Figure 18: Linkability vs. $C_{\text{llr}}^{\text{min}}$ results for the three attack models. Each point in the figure represents results on a dataset from the set of all 12 VoicePrivacy development and test datasets. Higher $C_{\text{llr}}^{\text{min}}$ and smaller *linkability* values correspond to better privacy.



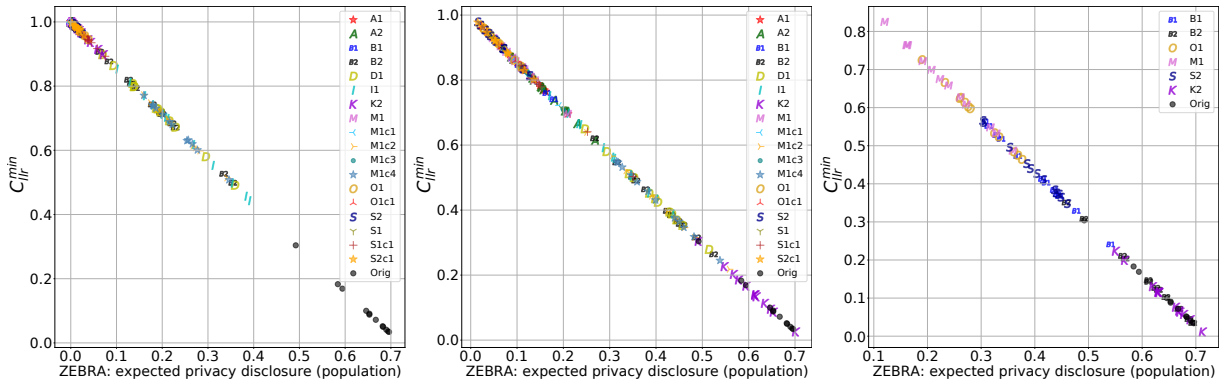
(a) Ignorant attacker (oa) (b) Lazy-informed attacker (aa) (c) Semi-informed attacker (aa)

Figure 19: EER vs. C_{llr} results for the three attack models. Each point in the figure represents results on a particular dataset from the set of all 12 VoicePrivacy development and test datasets.



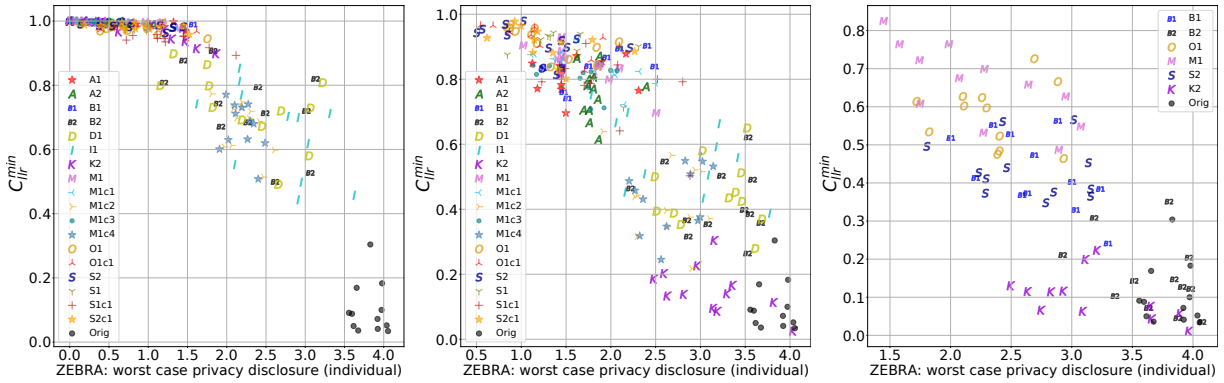
(a) Ignorant attacker (oa) (b) Lazy-informed attacker (aa) (c) Semi-informed attacker (aa)

Figure 20: EER vs. Linkability results for the three attack models. Each point in the figure represents results on a particular dataset from the set of all 12 VoicePrivacy development and test datasets.



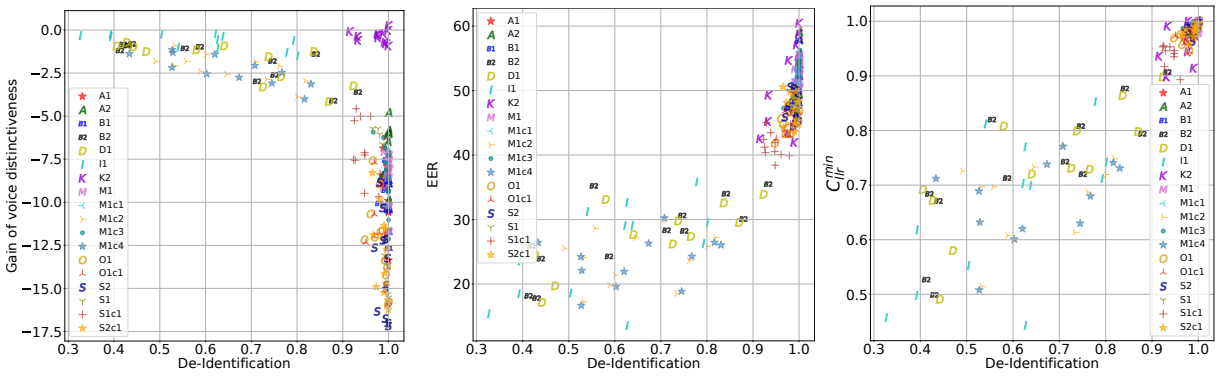
(a) Ignorant attacker (oa) (b) Semi-informed attacker (aa) (c) Semi-informed attacker (aa)

Figure 21: Expected privacy disclosure (population) vs. $C_{\text{llr}}^{\text{min}}$ results for the three attack models. Each point in the figure represents results on a particular dataset from the set of all 12 VoicePrivacy development and test datasets. Higher $C_{\text{llr}}^{\text{min}}$ and lower expected privacy disclosure metric values correspond to better privacy.



(a) Ignorant attacker (oa) (b) Semi-informed attacker (aa) (c) Semi-informed attacker (aa)

Figure 22: Worst case privacy disclosure (individual) vs. $C_{\text{llr}}^{\text{min}}$ results for the three attack models. Each point in the figure represents results on a particular dataset from the set of all 12 VoicePrivacy development and test datasets. Smaller worst case privacy disclosure metric values and higher $C_{\text{llr}}^{\text{min}}$ values correspond to better privacy.



(a) **DeID** vs. **gain of voice distinctiveness (G_{VD})** (b) **DeID** vs. **EER** (c) **DeID** vs. $\boldsymbol{C_{\text{llr}}^{\text{min}}}$

Figure 23: De-identification (DeID) vs. other metrics. Each point in the figure represents results on a particular dataset from the set of all 12 VoicePrivacy development and test datasets. In (b) and (c), the ignorant attack model is used to compute EER and $C_{\text{llr}}^{\text{min}}$. Higher DeID corresponds to better privacy.

4. Subjective evaluation

This section presents subjective evaluation results for speaker verifiability, speech naturalness, and speech intelligibility.

4.1. Subjective evaluation on verifiability, naturalness, and intelligibility

These three metrics are evaluated via a unified subjective evaluation test. The input speech trial can be an original or anonymized test set trial from a target or a non-target speaker. For intelligibility of the input trial, the evaluators assign a score from 1 ('totally unintelligible') to 10 ('totally intelligible'). For naturalness, the evaluators assign a score from 1 ('totally unnatural') to 10 ('totally natural'). For verifiability, the evaluators are required to listen to one original enrollment utterance of the target speaker and rate the similarity between the input trial and the enrollment voice using a scale of 1 to 10, where 1 denotes 'different speakers' and 10 denotes 'the same speaker' with highest confidence. The evaluators are instructed to assign the scores through a role-playing game.

Instructions for role-playing

When an evaluator started an evaluation session, the following instruction was displayed:[6]

> “Please imagine that you are working at a TV or radio company. You wish to broadcast interviews of person X, but this person X does not want to disclose his/her identity. Therefore you need to modify speech signals in order to hide it. You have several automated tools to change speaker identity. Some of them hide the identity well, but severely degrade audio quality. Some of them hide the identity, but the resulting speech sounds very unnatural and may become less intelligible. In such cases, the privacy of person X is protected, but you will receive many complaints from the audience and listeners of TV/radio programs. You need to balance privacy of person X and satisfaction of TV/radio program audience and listeners. Your task is to evaluate such automated tools to change speaker identity and find out well-balanced tools.”

Each of the three subjective metrics had a detailed instruction. The evaluator was asked to imagine the scene when evaluating the corresponding metric.

> “**Subjective speech naturalness**
> You will listen to either original audio and audio modified by the above anonymization tools. Some of them result in artifacts and degradation due to poor audio processing.
> Now, please listen to audio A and answer how much you can hear the audio degradation. Please judge based on the characteristics of the audio rather than what is being said.
> You need to select a score between 1 and 10, where a higher score indicates less degradation. In particular, 1 means “audio A exhibits severe audio degradation” and 10 means “audio A does not exhibit any degradation”. Please note that the original audio includes background noise.”

> “**Subjective speaker verifiability (similarity)**
> Your next task is to compare the processed or unprocessed audio A with audio B where the original person may speak different sentences. From the voices, you must determine whether they are from the same person or another person.
> Now, please listen to audio A above and audio B below, and determine if they were uttered by the same speaker. Please judge based on the characteristics of the voice rather than what is being said.
> You need to select one score between 1 and 10, where a higher score denotes higher speaker similarity. In particular, 1 means “audio A and B were uttered by different speakers for sure” and 10 means “audio A and B were uttered by the same speaker for sure.””

[6]The bank call center scenario mentioned in the evaluation plan was eventually replaced by this one.

> “
> **Subjective speech intelligibility**
> For the final task, you are required to listen to audio A again and try to understand the audio content. Please judge how understandable audio A is. You need to select one score between 1 and 10, where a higher score denotes higher intelligibility. In particular, 1 means “audio A is NOT understandable at all” and 10 means “audio A is perfectly understandable.”
> ”

4.2. Score distribution in violin plot

To reduce the perceptual bias of each individual evaluator, the naturalness, intelligibility, and verifiability scores collected via the unified subjective test were processed using normalized-rank normalization

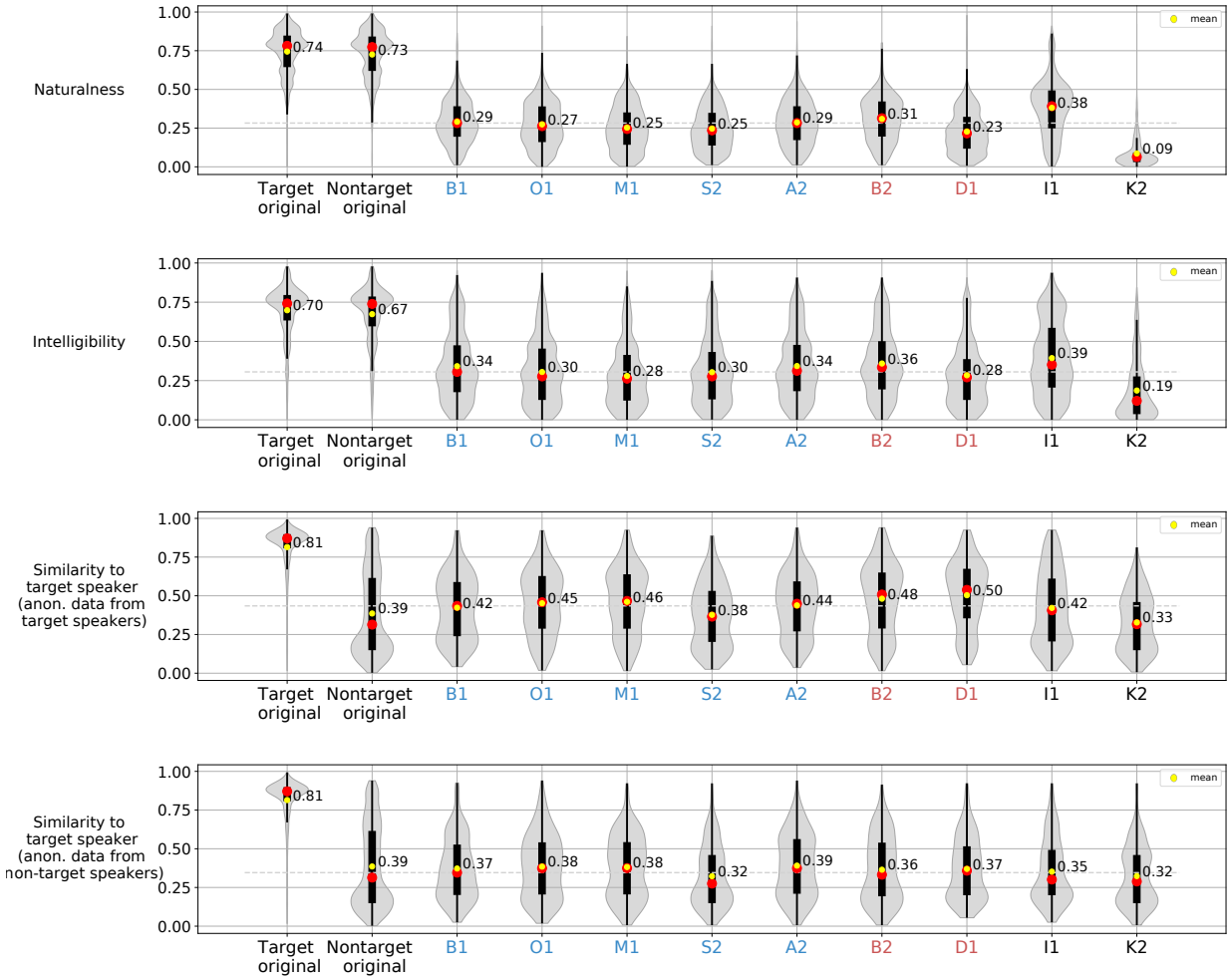


Figure 24: Violin plots of subjective speech naturalness, intelligibility, and speaker similarity obtained from the normalized scores. For naturalness and intelligibility, the scores for target and non-target anonymized data are pooled; for similarity, the scores for anonymized target and non-target speaker data are plotted separately in the 3rd and 4th sub-figures, respectively. The dotted line indicates the median of **B1**. Numbers correspond to mean values. Higher values for naturalness and intelligibility correspond to better utility, and lower scores for similarity to target speaker with anonymized data from target speaker indicate better privacy. Blue and red colors in the system notations indicate systems developed from **B1** and **B2**, respectively.

(Rosenberg & Ramabhadran, 2017). The processed scores are float numbers varying from 0 to 1. The Mann-Whiteney-U test was further used for statistical significance tests (Rosenberg & Ramabhadran, 2017).

The score distributions pooled over the three test sets are displayed in Figure 24 as violin plots (Hintze & Nelson, 1998). There are four types of trials: original or anonymized trials from target or non-target speakers. When displaying the results for naturalness and intelligibility, we merge the anonymized trials of both target and non-target speakers. When displaying the results for similarity, we keep them separate so that we can tell how well the speech of the target speakers has been anonymized. This is the reason why there are four sub-figures in Figure 24. Note that we display the scores of target and non-target original trials separately.

The significance test evaluates whether the differences between the scores of two systems are statistically significant. We followed Rosenberg & Ramabhadran (2017) and used the two-sided Mann-Whiteney test. The results are show in Tables 16 and 17. As the first row of Table 16a demonstrates, the scores of anonymized target speaker data from all the systems are statistically different from those of the original target speaker data. From the first two rows of Tables 17b and 17a, the results indicate that the scores of anonymized data from all the systems are statistically different from those of the original data.

Table 16: Significance test on subjective speaker similarity results pooled over *LibriSpeech-test, VCTK-test (common)*, and *VCTK-test (different)*. Cells in blue color denotes statistical significance ($p \ll 0.01$), while gray color denotes insignificant difference ($p > 0.01$). Blue and red colors in the system notations indicate systems developed from **B1** and **B2**, respectively. **Tar** and **Non-tar** denote natural speech from target and non-target speakers, respectively. Scores of anonymized target and non-target speakers data were separated, and the results are listed in (a) and (b) below, respectively.

(a) Anonymized speech of target speakers versus **Tar**, **Non-tar**

	Tar	**Non-tar**	**B1**	**O1**	**M1**	**S2**	**A2**	**B2**	**D1**	**I1**	**K2**
Tar		$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$
Non-tar	$\ll 0.01$		$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	0.510	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	0.0002	0.0019
B1	$\ll 0.01$	$\ll 0.01$		0.027	0.0057	0.0007	0.258	$\ll 0.01$	$\ll 0.01$	0.660	$\ll 0.01$
O1	$\ll 0.01$	$\ll 0.01$	0.027		0.483	$\ll 0.01$	0.266	0.042	0.0002	0.018	$\ll 0.01$
M1	$\ll 0.01$	$\ll 0.01$	0.0057	0.483		$\ll 0.01$	0.077	0.167	0.0030	0.0047	$\ll 0.01$
S2	$\ll 0.01$	0.510	0.0007	$\ll 0.01$	$\ll 0.01$		$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	0.0087	0.0004
A2	$\ll 0.01$	$\ll 0.01$	0.258	0.266	0.077	$\ll 0.01$		0.0020	$\ll 0.01$	0.135	$\ll 0.01$
B2	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	0.042	0.167	$\ll 0.01$	0.0020		0.140	0.0001	$\ll 0.01$
D1	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	0.0002	0.0030	$\ll 0.01$	$\ll 0.01$	0.140		$\ll 0.01$	$\ll 0.01$
I1	$\ll 0.01$	0.0002	0.660	0.018	0.0047	0.0087	0.135	0.0001	$\ll 0.01$		$\ll 0.01$
K2	$\ll 0.01$	0.0019	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	0.0004	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	

(b) Anonymized speech of non-target speakers versus **Tar**, **Non-tar**

	Tar	**Non-tar**	**B1**	**O1**	**M1**	**S2**	**A2**	**B2**	**D1**	**I1**	**K2**
Tar		$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$
Non-tar	$\ll 0.01$		0.440	0.171	0.241	0.0004	0.038	0.653	0.634	0.411	0.0005
B1	$\ll 0.01$	0.440		0.327	0.379	0.0002	0.119	0.490	0.839	0.102	0.0003
O1	$\ll 0.01$	0.171	0.327		0.943	$\ll 0.01$	0.546	0.122	0.267	0.012	$\ll 0.01$
M1	$\ll 0.01$	0.241	0.379	0.943		$\ll 0.01$	0.478	0.167	0.293	0.018	$\ll 0.01$
S2	$\ll 0.01$	0.0004	0.0002	$\ll 0.01$	$\ll 0.01$		$\ll 0.01$	0.0047	0.0004	0.033	0.977
A2	$\ll 0.01$	0.038	0.119	0.546	0.478	$\ll 0.01$		0.036	0.078	0.0016	$\ll 0.01$
B2	$\ll 0.01$	0.653	0.490	0.122	0.167	0.0047	0.036		0.603	0.467	0.0066
D1	$\ll 0.01$	0.634	0.839	0.267	0.293	0.0004	0.078	0.603		0.187	0.0006
I1	$\ll 0.01$	0.411	0.102	0.012	0.018	0.033	0.0016	0.467	0.187		0.038
K2	$\ll 0.01$	0.0005	0.0003	$\ll 0.01$	$\ll 0.01$	0.977	$\ll 0.01$	0.0066	0.0006	0.038	

Table 17: Significance test on subjective naturalness and intelligibility results pooled over *LibriSpeech-test, VCTK-test (common)*, and *VCTK-test (different)*. Cells in blue color denotes statistical significance ($p \ll 0.01$), while gray color denotes ($p > 0.01$). **Tar** and **Non-tar** denote natural speech from target and non-target speakers, respectively. Blue and red colors in the system notations indicate systems developed from **B1** and **B2**, respectively. Scores of both anonymized target and non-target data were merged for each system.

(a) Naturalness

	Tar	**Non-tar**	**B1**	**O1**	**M1**	**S2**	**A2**	**B2**	**D1**	**I1**	**K2**
Tar		$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$
Non-tar	$\ll 0.01$		$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$
B1	$\ll 0.01$	$\ll 0.01$		0.017	$\ll 0.01$	$\ll 0.01$	0.562	0.0027	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$
O1	$\ll 0.01$	$\ll 0.01$	0.017		0.0019	$\ll 0.01$	0.081	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$
M1	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	0.0019		0.398	$\ll 0.01$	$\ll 0.01$	0.0003	$\ll 0.01$	$\ll 0.01$
S2	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	0.398		$\ll 0.01$	$\ll 0.01$	0.0026	$\ll 0.01$	$\ll 0.01$
A2	$\ll 0.01$	$\ll 0.01$	0.562	0.081	$\ll 0.01$	$\ll 0.01$		0.0006	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$
B2	$\ll 0.01$	$\ll 0.01$	0.0027	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	0.0006		$\ll 0.01$	$\ll 0.01$	$\ll 0.01$
D1	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	0.0003	0.0026	$\ll 0.01$	$\ll 0.01$		$\ll 0.01$	$\ll 0.01$
I1	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$		$\ll 0.01$
K2	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	

(b) Intelligibility

	Tar	**Non-tar**	**B1**	**O1**	**M1**	**S2**	**A2**	**B2**	**D1**	**I1**	**K2**
Tar		$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$
Non-tar	$\ll 0.01$		$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$
B1	$\ll 0.01$	$\ll 0.01$		0.0003	$\ll 0.01$	$\ll 0.01$	0.866	0.043	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$
O1	$\ll 0.01$	$\ll 0.01$	0.0003		0.0053	0.764	0.0001	$\ll 0.01$	0.048	$\ll 0.01$	$\ll 0.01$
M1	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	0.0053		0.013	$\ll 0.01$	$\ll 0.01$	0.421	$\ll 0.01$	$\ll 0.01$
S2	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	0.764	0.013		$\ll 0.01$	$\ll 0.01$	0.101	$\ll 0.01$	$\ll 0.01$
A2	$\ll 0.01$	$\ll 0.01$	0.866	0.0001	$\ll 0.01$	$\ll 0.01$		0.064	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$
B2	$\ll 0.01$	$\ll 0.01$	0.043	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	0.064		$\ll 0.01$	0.0059	$\ll 0.01$
D1	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	0.048	0.421	0.101	$\ll 0.01$	$\ll 0.01$		$\ll 0.01$	$\ll 0.01$
I1	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	0.0059	$\ll 0.01$		$\ll 0.01$
K2	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	$\ll 0.01$	

4.3. DET curves

To investigate the difference across systems quantitatively, we compute detection error trade-off (DET) curves (Martin et al., 1997) based on the score distribution. Since there are four types of scores, i.e., {target original, non-target original, target anonymized, non-target anonymized}, we computed the DET curves in the following ways:

- Naturalness and intelligibility DET curves: the positive class (anchor) is "target original", and the negative class is either "non-target original" or "anonymized (target and non-target)" from one anonymization system;
- Similarity DET curve 1: the positive class is "target original" or "target anonymized" from one anonymization system, and the negative class (anchors) is "non-target original";
- Similarity DET curve 2: the positive class is "target original" or "non-target anonymized" from one anonymization system, and the negative class (anchors) is "non-target original".

For naturalness and intelligibility, an ideal anonymization system should have a DET curve close to that of original data, indicating similar naturalness and intelligibility scores to the original data and therefore

minimum degradation on naturalness and intelligibility. For similarity curve 1, an ideal anonymization system should have a DET curve close to the diagonal line from bottom-right to top-left, indicating that the anonymized data of a target speaker sounds similar to the non-target data.

The four types of DET curves are plotted in Figure 25. As the top two sub-figures demonstrate, the DET curves of the original data are straight lines across the (50%, 50%) point, indicating that the scores of non-target original data are similar to those of the target original data. This is expected because original data should have similar naturalness and intelligibility no matter whether they are from target or non-target speakers. In contrast, the DET curves of anonymized systems are not close to the curve of original data, suggesting that anonymized data are inferior to the original data in terms of naturalness and intelligibility, similar to the messages from the violin plot in previous section.

Among the anonymized systems, the naturalness DET curves of **I1** and **K2** seem to deviate from other systems. While other systems are based on either **B1** or **B2**, **I1** uses a different signal-processing-based approach to change the speech spectra, and **K2** uses a different deep learning method. **I1**'s framework avoids several types of errors such as speech recognition in **B1**, which may contribute to its performance. However, it is interesting to note how different signal processing algorithms result in different perceptual naturalness and intelligibility. Also note that none of the systems except **I1** outperformed **B2**.

Concerning similarity, let us focus on the case where target speaker data is anonymized (left-bottom figure in 25). We observe that the DET curve of original data is close to the bottom-left corner while those of anonymized data are close to top-right corner. In other words, the anonymized target speaker

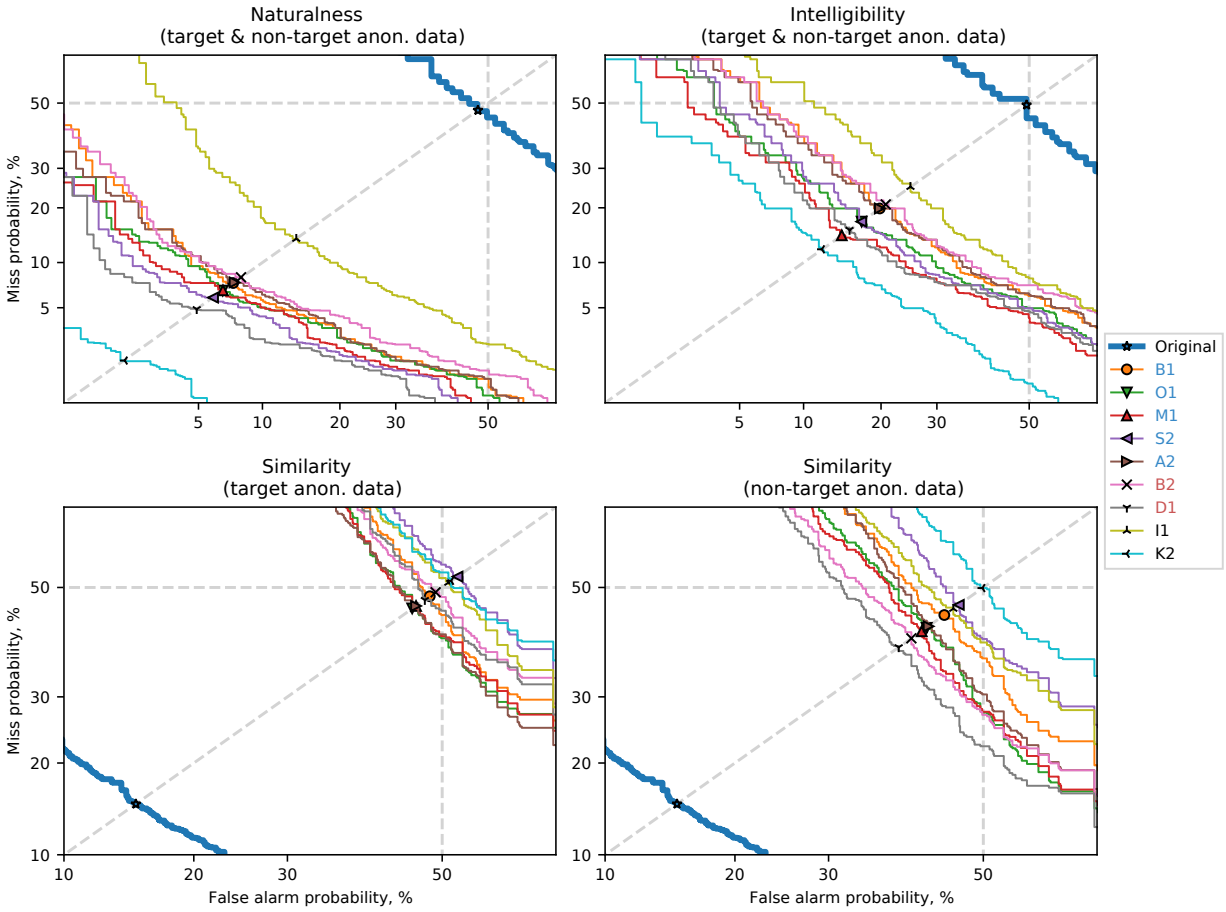
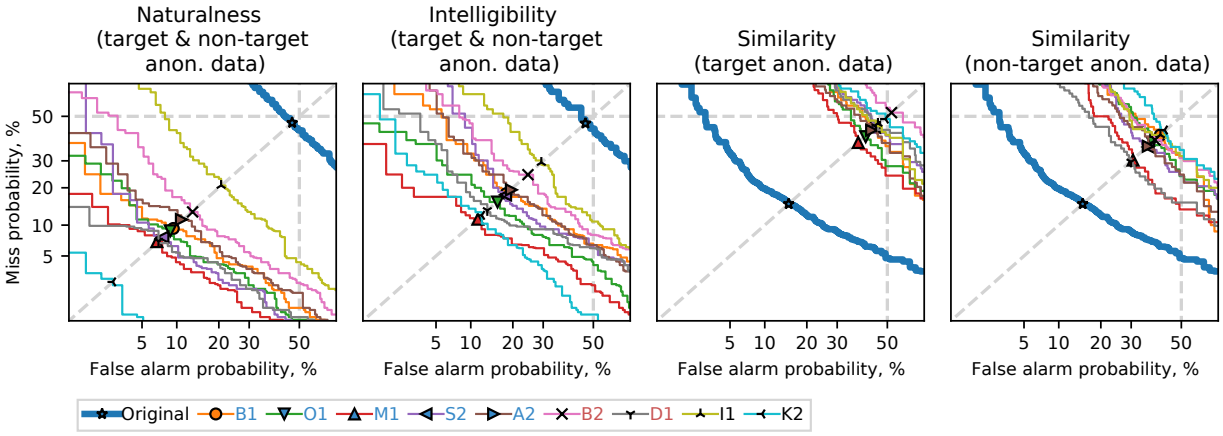
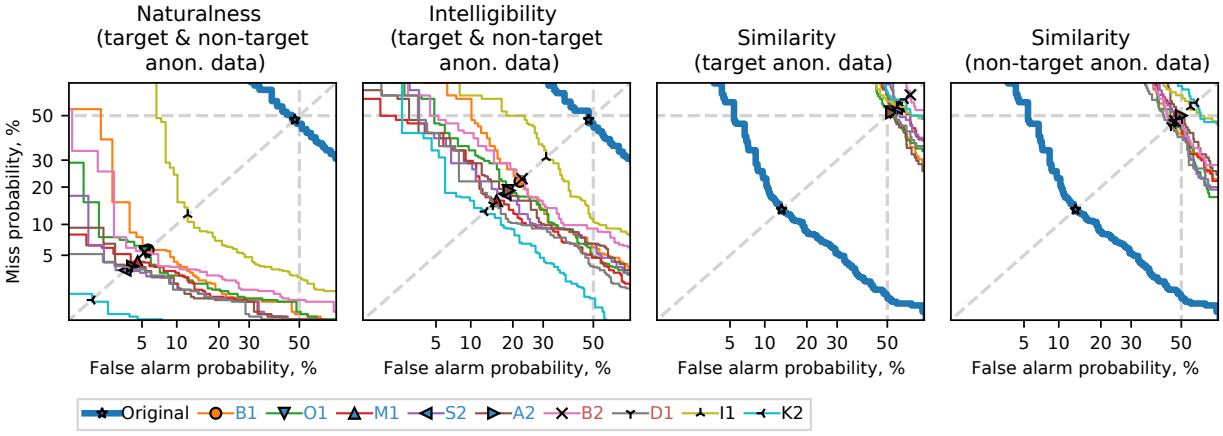


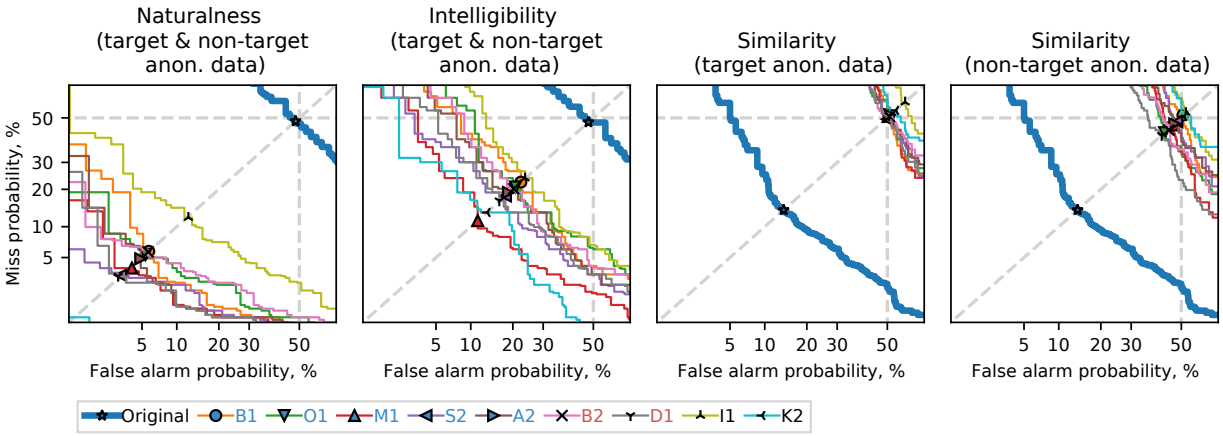
Figure 25: DET curves based on subjective evaluation scores pooled over *LibriSpeech-test* and *VCTK-test* datasets.



(a) LibriSpeech-test



(b) VCTK-test (different)



(c) VCTK-test (common)

Figure 26: DET curves for each test set.

data produced similar perceptual scores to the non-target speaker data, indicating that anonymized target speaker data sound less similar to the original target speaker. Similar results can be observed from the curves in Figure 26, which are separately plotted on the three test sets.

The similarity DET curves of **K2**, **S2**, and **I1** on target speaker data seem to be closer to the (50%, 50%) point than others (left-bottom sub-figure in Figure 25). However, the three systems are quite different in terms of naturalness and intelligibility, particularly with **I1** and **K2** achieving the highest and lowest median MOS result, respectively. This implies that an anonymized trial may sound like the voice of a different speaker simply because of the severe distortion caused by anonymization.

In summary, all the submitted anonymization systems can anonymize the perceived speaker identity to some degree. However, none of them can produce an anonymized trial that is as natural and intelligible as original speech data. One signal-processing-based anonymization method may degrade the naturalness and intelligibility of anonymized trials less severely, but still introduces some degradation in that regard.

5. Comparison of objective and subjective evaluation results

In this section, we are interested in comparing objective and subjective evaluation results. From the subjective speaker verifiability (similarity) scores, we computed EER, ROCCH-EER, C_{llr}, and $C_{\text{llr}}^{\text{min}}$. We then compare these metrics with those obtained from objective speaker verifiability scores. These comparisons are plotted in Figures 27, 28, 29, and 30.

The marker "Enr: o, Trl: o" in Figure 27 denotes original trials, and other markers denote anonymized trials from the submitted systems. The comparison between original and anonymized trials indicates that both objective and subjective EERs increase after the trial is anonymized. However, the increase varies across the anonymization systems and test sets. Similar results can be observed for ROCCH-EER, C_{cllr}, and $C_{\text{cllr}}^{\text{min}}$. Furthermore, the objective and subjective EERs are positively correlated (Figures 28, 29 and 30). This suggests that the concerned anonymization methods can hide the speaker identity to some degree from both ASV system and human ears. This is an encouraging message from the challenge.

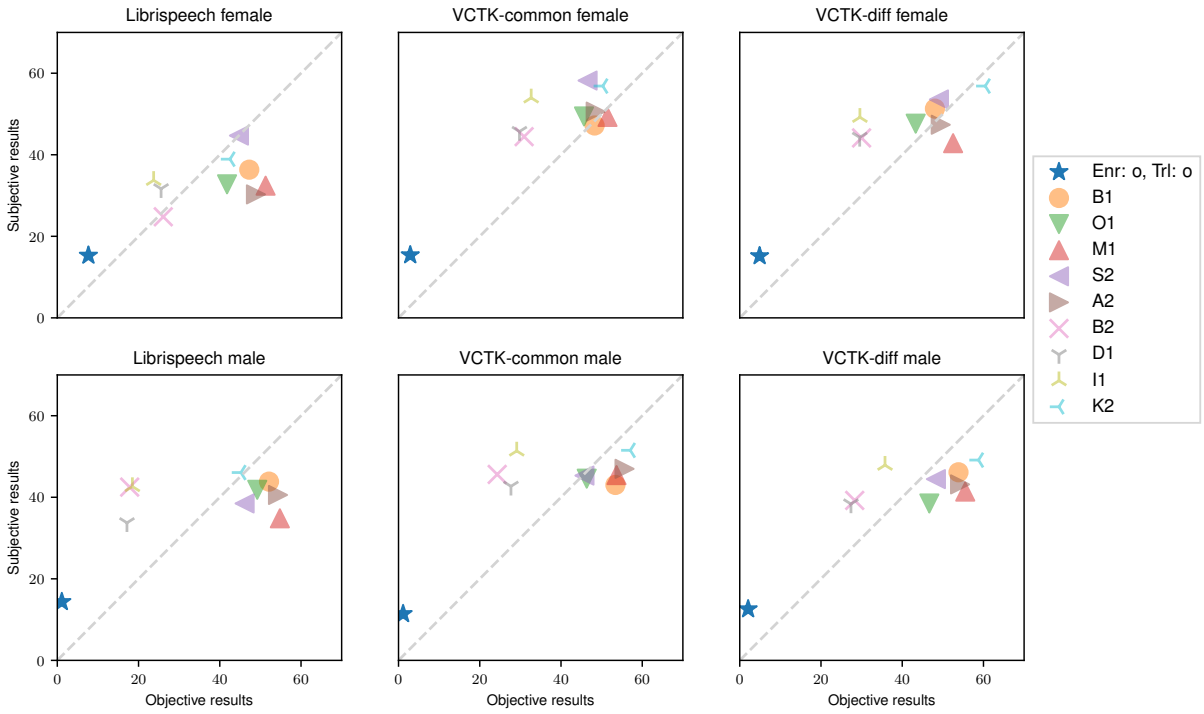


Figure 27: Objective versus subjective EER.

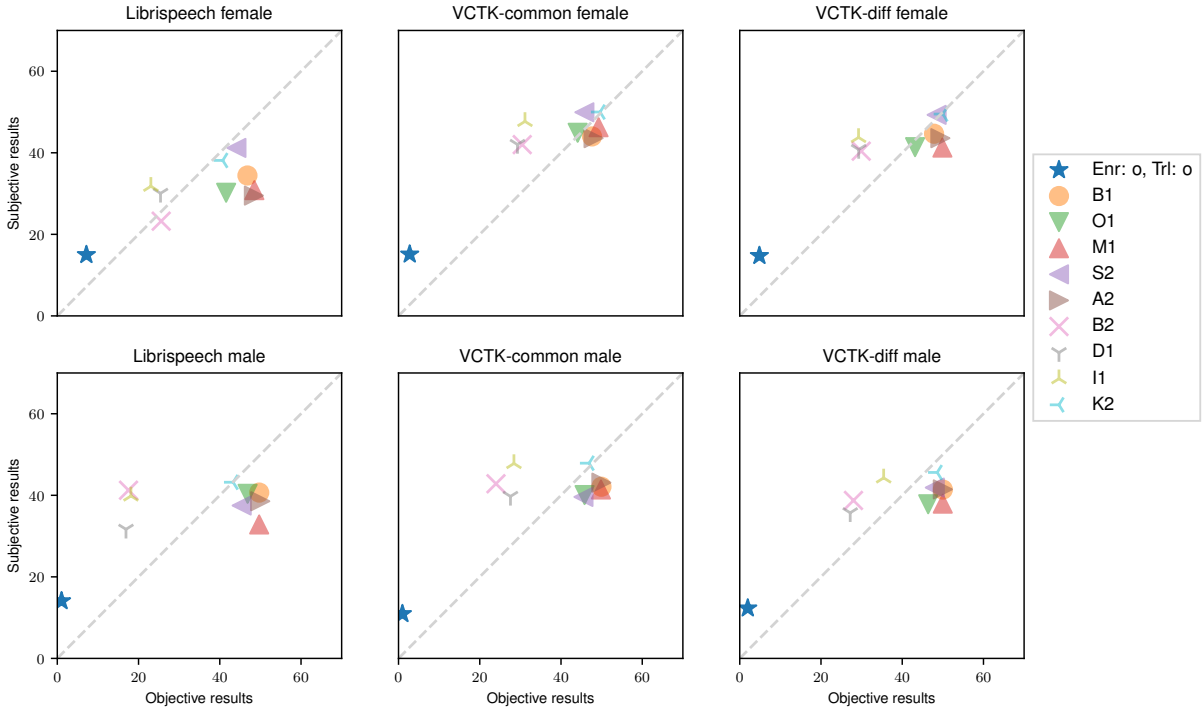


Figure 28: Objective versus subjective ROCCH-EER.

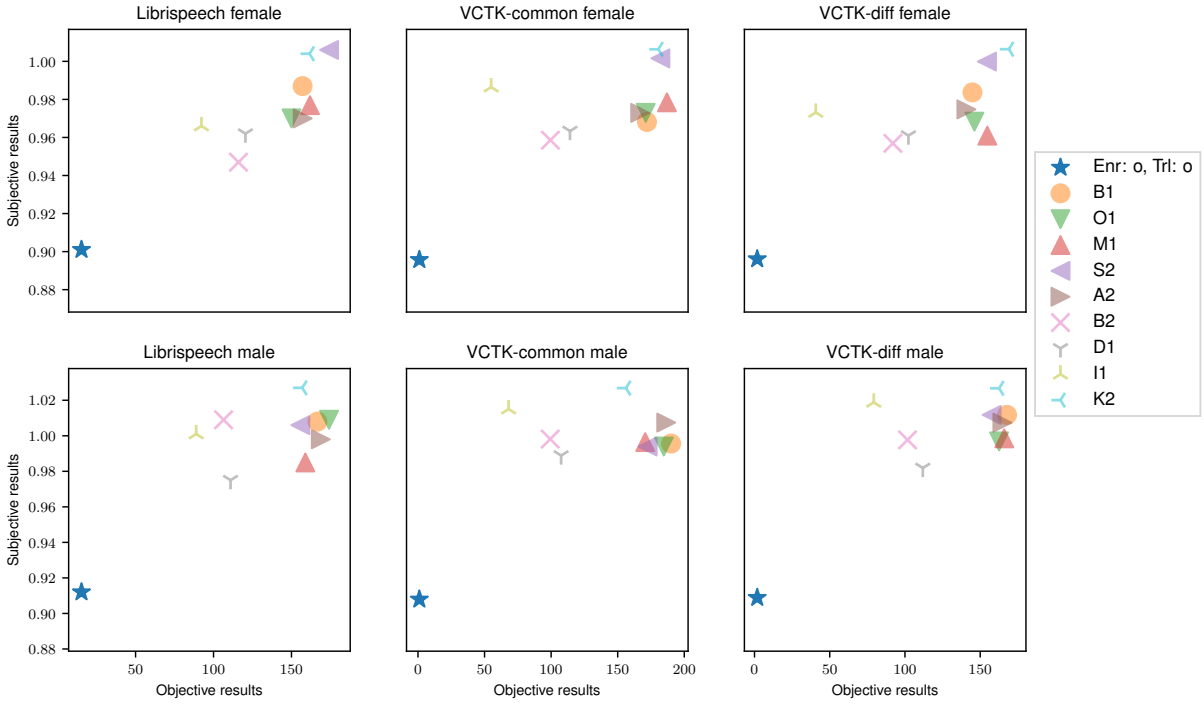


Figure 29: Objective versus subjective C_{llr}.

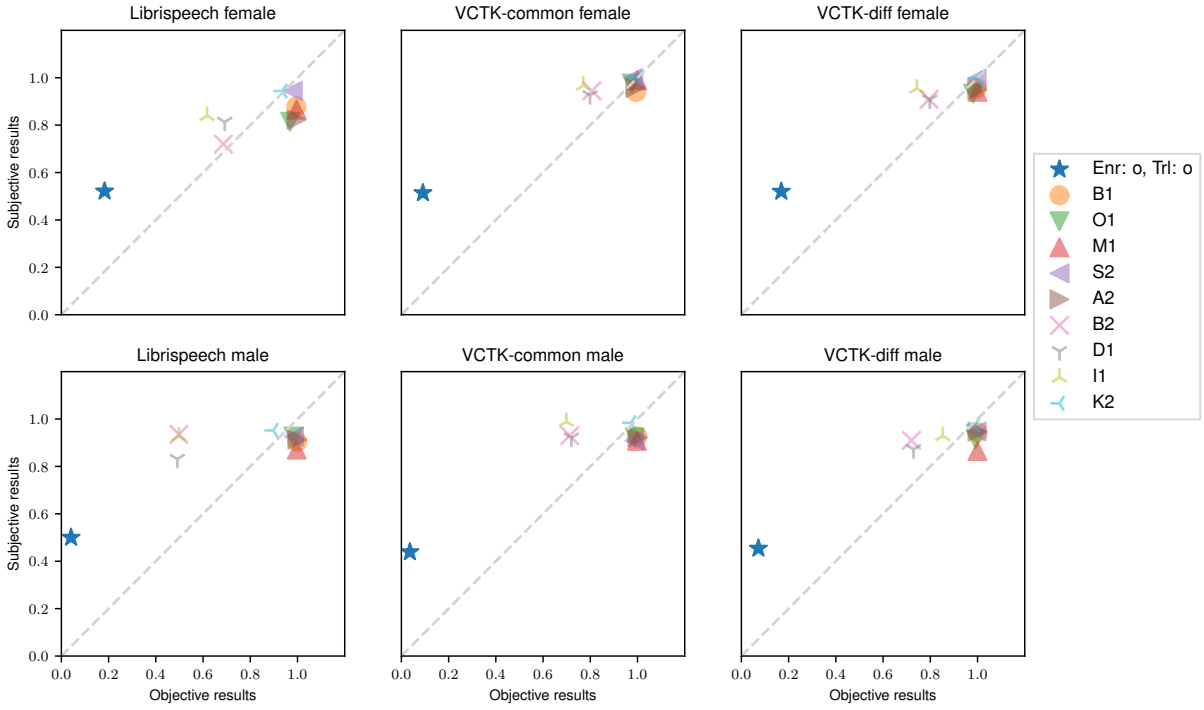


Figure 30: Objective versus subjective $C_{\text{llr}}^{\text{min}}$.

References

Brummer, N. (2010). *Measuring, refining and calibrating speaker and language information extracted from speech*. Ph.D. thesis Stellenbosch: University of Stellenbosch.

Brümmer, N., & De Villiers, E. (2011). The BOSARIS toolkit user guide: Theory, algorithms and code for binary classifier score processing.

Brümmer, N., & Du Preez, J. (2006). Application-independent evaluation of speaker detection. *Computer Speech and Language*, *20*, 230–275.

Champion, P., Jouvet, D., & Larcher, A. (2020). *Speaker information modification in the VoicePrivacy 2020 toolchain*. Research Report INRIA Nancy, équipe Multispeech ; LIUM - Laboratoire d'Informatique de l'Université du Mans. URL: `https://hal.archives-ouvertes.fr/hal-02995855`.

Dubagunta, S. P., van Son, R. J., & Doss, M. M. (2020). Adjustable deterministic pseudonymisation of speech: Idiap-NKI's submission to VoicePrivacy 2020 challenge, . URL: `https://www.voiceprivacychallenge.org/docs/Idiap-NKI.pdf`.

Espinoza-Cuadros, F. M., Perero-Codosero, J. M., Antón-Martín, J., & Hernández-Gómez, L. A. (2020). Speaker de-identification system using autoencoders and adversarial training, . `arXiv:2011.04696`.

Gomez-Barrero, M., Galbally, J., Rathgeb, C., & Busch, C. (2017). General framework to evaluate unlinkability in biometric template protection systems. *IEEE Transactions on Information Forensics and Security*, *13*, 1406–1420.

Gupta, P., Prajapati, G. P., Singh, S., Kamble, M. R., & Patil, H. A. (2020). Design of voice privacy system using linear prediction. URL: `https://www.voiceprivacychallenge.org/docs/DA-IICT-Speech-Group.pdf`.

Han, Y., Li, S., Cao, Y., & Yoshikawa, M. (2020). System description for Voice Privacy Challenge. Kyoto team, . URL: `https://www.voiceprivacychallenge.org/docs/Kyoto.pdf`.

Hintze, J. L., & Nelson, R. D. (1998). Violin plots: a box plot-density trace synergism. *The American Statistician*, *52*, 181–184.

Huang, C.-L. (2020). Analysis of PingAn submission in the VoicePrivacy 2020 Challenge. URL: `https://www.voiceprivacychallenge.org/docs/PingAn.pdf`.

Maouche, M., Srivastava, B. M. L., Vauquier, N., Bellet, A., Tommasi, M., & Vincent, E. (2020). A comparative study of speech anonymization metrics. In *Interspeech* (pp. 1708–1712).

Martin, A., Doddington, G., Kamm, T., Ordowski, M., & Przybocki, M. (1997). *The DET curve in assessment of detection task performance*. Technical Report National Inst of Standards and Technology Gaithersburg MD.

Mawalim, C. O., Galajit, K., Karnjana, J., & Unoki, M. (2020). X-vector singular value modification and statistical-based decomposition with ensemble regression modeling for speaker anonymization system. In *Interspeech* (pp. 1703–1707).

Nautsch, A., Patino, J., Tomashenko, N., Yamagishi, J., Noé, P.-G., Bonastre, J.-F., Todisco, M., & Evans, N. (2020). The privacy ZEBRA: Zero evidence biometric recognition assessment. In *Interspeech* (pp. 1698–1702).

Noé, P.-G., Bonastre, J.-F., Matrouf, D., Tomashenko, N., Nautsch, A., & Evans, N. (2020). Speech pseudonymisation assessment using voice similarity matrices. In *Interspeech* (pp. 1718–1722).

Patino, J., Tomashenko, N., Todisco, M., Nautsch, A., & Evans, N. (2021). Speaker anonymisation using the McAdams coefficient. In *Interspeech* (pp. 1099–1103).

Provost, F., & Fawcett, T. (2001). Robust classification for imprecise environments. *Machine Learning*, *42*, 203–231.

Qian, J., Han, F., Hou, J., Zhang, C., Wang, Y., & Li, X.-Y. (2018). Towards privacy-preserving speech data publishing. In *2018 IEEE Conference on Computer Communications (INFOCOM)* (pp. 1079–1087).

Ramos, D., & Gonzalez-Rodriguez, J. (2008). Cross-entropy analysis of the information in forensic speaker recognition. In *Odyssey*.

Rosenberg, A., & Ramabhadran, B. (2017). Bias and statistical significance in evaluating speech synthesis with mean opinion scores. In *Interspeech* (pp. 3976–3980).

Srivastava, B. M. L., Tomashenko, N., Wang, X., Vincent, E., Yamagishi, J., Maouche, M., Bellet, A., & Tommasi, M. (2020a). Design choices for x-vector based speaker anonymization. In *Interspeech* (pp. 1713–1717).

Srivastava, B. M. L., Vauquier, N., Sahidullah, M., Bellet, A., Tommasi, M., & Vincent, E. (2020b). Evaluating voice conversion-based privacy protection against informed attackers. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2802–2806).

Tomashenko, N., Srivastava, B. M. L., Wang, X., Vincent, E., Nautsch, A., Yamagishi, J., Evans, N., Patino, J., Bonastre, J.-F., Noé, P.-G., & Todisco, M. (2020a). The VoicePrivacy 2020 Challenge evaluation plan, . URL: `https://www.voiceprivacychallenge.org/docs/VoicePrivacy_2020_Eval_Plan_v1_3.pdf`.

Tomashenko, N., Srivastava, B. M. L., Wang, X., Vincent, E., Nautsch, A., Yamagishi, J., Evans, N., Patino, J., Bonastre, J.-F., Noé, P.-G., & Todisco, M. (2020b). Introducing the VoicePrivacy initiative. In *Interspeech* (pp. 1693–1697).

Tomashenko, N., Wang, X., Vincent, E., Patino, J., Srivastava, B. M. L., Noé, P.-G., Nautsch, A., Evans, N., Yamagishi, J., O'Brien, B., Chanclu, A., Bonastre, J.-F., Todisco, M., & Maouche, M. (2021). The VoicePrivacy 2020 Challenge: Results and findings. `arXiv:2109.00648` submitted to Special Issue on Voice Privacy in Computer Speech and Language.

Turner, H., Lovisotto, G., & Martinovic, I. (2020). Speaker anonymization with distribution-preserving x-vector generation for the VoicePrivacy Challenge 2020, . `arXiv:2010.13457`.